2021 AMIA Board of Directors

Officers

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

Chair/President

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

Chair/President-Elect

Neil Sarkar, PhD, MLIS, FACMI
Brown University

Treasurer

Theresa A. Cullen, MD, FAMIA
Regenstrief Institute

Secretary

Directors

Julia Adler-Milstein, PhD, FACMI
University of California, San Francisco

Tiffany J. Bright, PhD, FACMI
IBM Watson Health

William Brown III, PhD, DrPH, MA
University of California, San Francisco

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

Susan C. Hull, MSN, RN-BC, NEA-BC, FAMIA
MITRE Corporation

Laura K. Heerman Langford, PhD, RN, FAMIA
Intermountain Healthcare

Philip R.O. Payne, PhD, FACMI, FAMIA
Washington University in St. Louis

Wanda Pratt, PhD, FACMI
University of Washington

S. Trent Rosenbloom, MD, MPH, FACMI
Vanderbilt University Medical Center
Victoria L. Tiase, PhD, RN-BC, FAMIA, FAAN
Working Group Steering Committee Representative
NewYork-Presbyterian Hospital

Adam Wright, PhD, FACMI, FAMIA, FIAHSI
Vanderbilt University Medical Center

Li Zhou, MD, PhD, FACMI, FAMIA
Brigham and Women’s Hospital/Harvard Medical School

Ex-Officio Board Members

Genevieve Melton-Meaux, MD, PhD, FACMI
American College of Medical Informatics President
University of Minnesota

Josette Jones, RN
Academic Forum Executive Committee Chair
Indiana University

Eileen Koski, MPhil, FAMIA
Informatics Partnership Council Chair
IBM Research

Jonathan R. Nebeker, MD, MS
Health Systems Chair
Veterans’ Health Administration

Lynda Hoeksema, DNP, FNP-BC, RN-BC, FAMIA
Student Working Group Representative
Ohio State University

Tanya Tolpegin, MBA, CAE
Chief Executive Officer
AMIA
AMIA 2021 Annual Symposium Scientific Program Committee

Chair
Adam B. Wilcox, PhD, FACMI
Washington University in St. Louis

Vice Chairs
Randi Foraker, PhD, MA, FAHA, FAMIA
Washington University in St. Louis

Kensaku Kawamoto, MD, PhD, MHS, FACMI, FAMIA
University of Utah

Yves A. Lussier, MD, FACMI
University of Utah

Nadine McCleary, MD, MPH
Dana-Farber Cancer Institute

Members
Joanna Abraham, PhD
Washington University in St. Louis

Paul Aiyetan, MD, MS, PhD
Frederick National Laboratory for Cancer Research

Suzanne Boren, MHA, PhD
University of Missouri

Andrew Boyd, MD
University of Illinois at Chicago

Thomas Campion, PhD
Weill Cornell Medicine

Jin Chen, PhD
University of Kentucky

Yunan Chen, PhD
University of California, Irvine

Amar Das, MD, PhD
Merck
Karen Dunn Lopez, PhD, MPH, RN
The University of Florida

Jose Florez-Arango, MD, PhD, MS
Texas A&M University

Kate Fultz Hollis, MS, MBI
Oregon Health & Science University

Jennifer Garvin, PhD, MBA
The Ohio State University/VA

Tamara Goncalves Rezende Macieira, PhD, BSN
University of Florida

Assaf Gottlieb, PhD
University of Texas Health science center at Houston

Ilana Graetz, PhD
Emory Rollins School of Public Health

Eric Hall, PhD
Geisinger

Elizabeth Heitkemper, PhD, RN
The University of Texas at Austin

Harry Hochheiser, PhD
University of Pittsburgh

Felix Holl, MSc, PhD (candidate)
Neu-Ulm University of Applied Sciences/University of Munich

Julian Hong, MD, MS
University of California, San Francisco

Robert Jenders, MD, MS
University of California, Los Angeles

Suranga Kasthurirathne, PhD
Indiana University School of Medicine/Regenstrief Institute

Halil Kilicoglu, PhD
University of Illinois at Urbana-Champaign

Fabrício Kury, MD
Regeneron Genetics Center
Christoph Lehmann, MD  
UT Southwestern

Bruce Levy, MD  
Geisinger

Carol Macumber, MS, PMP, FAMIA  
Clinical Architecture

Ruth Masterson Creber, PhD  
Weill Cornell Medicine

Mollie McKillop, PhD, MPH  
IBM Watson Health

Xia Ning, PhD  
The Ohio State University Wexner Medical Center

Carlos Luis Parra-Calderón, MSc  
Andalusian Health Service - Virgen del Rocío University Hospital - Institute of Biomedicine of Seville

Eric Poon, MD, MPH  
Duke University Health System

Simon Poon, BSc, ME, MPH, PHD  
The University of Sydney

Fabian Prasser, Prof. Dr.  
Charité, Universitätsmedizin Berlin

Arvind Rao, PhD  
University of Michigan

Blaine Reeder, PhD  
University of Missouri

Renato Marcos Sabbatini, PhD  
HL7 Institute Brazil

Hojjat Salmasian, MD, MPH, PhD  
Brigham and Women's Hospital / Harvard Medical School

Matthew Scotch, PhD, MPH  
Arizona State University

Jane Snowdon, PhD  
IBM Research
Gillian Strudwick, RN, BNSc, MN, PhD
Centre for Addiction and Mental Health/University of Toronto, Canada

Vignesh Subbian, PhD
University of Arizona

Saanie Sulley, MD, PhD
National Healthy Start Association

Carolyn Sun, PhD, RN, ANP-BC
Hunter-Bellevue School of Nursing

Jason Thomas, BS
University of Washington

Thankam Thyvalikakath, DMD, MDS, PhD
Regenstrief Institute and Indiana University School of Dentistry

Jill Tiongco, MD, FAMIA
Montage Health

Umit Topaloglu, PhD
Wake Forest School of Medicine

Ece Uzun, MS, PhD
Brown University

Kavishwar Waghlikar, MD PhD
Harvard Medical School

Theresa Walunas, PhD
Northwestern University

Keith Woeltje, MD, PhD
Washington University School of Medicine

Po-Yin Yen, PhD
Washington University in Saint Louis

Kun-Hsing Yu, MD, PhD
Harvard Medical School

Teresa Zayas Cabán, PhD
Office of the National Coordinator for Health Information Technology

Ping Zhang, PhD
The Ohio State University
<table>
<thead>
<tr>
<th>AMIA 2021 Annual Symposium Reviewers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aakre, Christopher</td>
</tr>
<tr>
<td>Aarts, Jos</td>
</tr>
<tr>
<td>Aastha, -</td>
</tr>
<tr>
<td>Abbott, Amy</td>
</tr>
<tr>
<td>Abedian, Sajjad</td>
</tr>
<tr>
<td>Abeysinghe, Rashmie</td>
</tr>
<tr>
<td>Abraham, Joanna</td>
</tr>
<tr>
<td>Abrams, Meredith</td>
</tr>
<tr>
<td>Achuthan, Srisairam</td>
</tr>
<tr>
<td>Adams, Griffin</td>
</tr>
<tr>
<td>Adapa, Karthik</td>
</tr>
<tr>
<td>Adejare Jr., Adeboye Abiola</td>
</tr>
<tr>
<td>Adekkanattu, Prakash</td>
</tr>
<tr>
<td>Adib, Riddhiman</td>
</tr>
<tr>
<td>Adrianto, Indra</td>
</tr>
<tr>
<td>Afshar, Majid</td>
</tr>
<tr>
<td>Agurto, Carla</td>
</tr>
<tr>
<td>Ahalt, Stanley</td>
</tr>
<tr>
<td>Ahmad, Suzan</td>
</tr>
<tr>
<td>Ahmad, Faraz</td>
</tr>
<tr>
<td>Ajjarapu, Samuel</td>
</tr>
<tr>
<td>Akbar, Saba</td>
</tr>
<tr>
<td>Akbilgic, Oguz</td>
</tr>
<tr>
<td>Name</td>
</tr>
<tr>
<td>--------------------------------</td>
</tr>
<tr>
<td>Avramovic, Sanja</td>
</tr>
<tr>
<td>Avrunin, George S</td>
</tr>
<tr>
<td>Babbrah, Pooja</td>
</tr>
<tr>
<td>Bacher, Ian</td>
</tr>
<tr>
<td>Bajracharya, Adarsha</td>
</tr>
<tr>
<td>Bakal, Mehmet Gokhan</td>
</tr>
<tr>
<td>Baker, Rebecca</td>
</tr>
<tr>
<td>Balatsoukas, Panagiotis</td>
</tr>
<tr>
<td>Baldi, Ileana</td>
</tr>
<tr>
<td>Balyan, Renu</td>
</tr>
<tr>
<td>Banach, Mary</td>
</tr>
<tr>
<td>Bandiera-Paiva, Paulo</td>
</tr>
<tr>
<td>Banerjee, Imon</td>
</tr>
<tr>
<td>Banerjee, Dipanjan</td>
</tr>
<tr>
<td>Bangash, Hana</td>
</tr>
<tr>
<td>Banning, Pamela</td>
</tr>
<tr>
<td>Barman, Arko</td>
</tr>
<tr>
<td>Barnhill, Rick</td>
</tr>
<tr>
<td>Barone, Eleanor Jane</td>
</tr>
<tr>
<td>Barrett, Laura</td>
</tr>
<tr>
<td>Barretto, Eluizio Henrique Saraiva</td>
</tr>
<tr>
<td>Bar-Shain, David</td>
</tr>
<tr>
<td>Bashir, Rabia</td>
</tr>
<tr>
<td>Bashir, Ayisha Zaka</td>
</tr>
<tr>
<td>Name</td>
</tr>
<tr>
<td>-----------------------------</td>
</tr>
<tr>
<td>Eisman, Aaron</td>
</tr>
<tr>
<td>El-Azab, Sarah</td>
</tr>
<tr>
<td>Eldredge, Christina</td>
</tr>
<tr>
<td>Elghohari, Baher</td>
</tr>
<tr>
<td>El-Kareh, Robert</td>
</tr>
<tr>
<td>Elmarakeby, Haitham</td>
</tr>
<tr>
<td>Elshehaly, Mai</td>
</tr>
<tr>
<td>Emerson, Beth</td>
</tr>
<tr>
<td>Epperlein, Jonathan</td>
</tr>
<tr>
<td>Epps, Mika</td>
</tr>
<tr>
<td>Eran, Alal</td>
</tr>
<tr>
<td>Eschrich, Steven</td>
</tr>
<tr>
<td>Estiri, Hossein</td>
</tr>
<tr>
<td>Etingen, Bella</td>
</tr>
<tr>
<td>Etu, Egbe-Etu Emmanuel</td>
</tr>
<tr>
<td>Everson, Jordan</td>
</tr>
<tr>
<td>Exposito, Ernesto</td>
</tr>
<tr>
<td>Eyigoz, Elif</td>
</tr>
<tr>
<td>Eyre, Hannah</td>
</tr>
<tr>
<td>Fabian, Lacy</td>
</tr>
<tr>
<td>Facelli, Julio</td>
</tr>
<tr>
<td>Fadden, Michael</td>
</tr>
<tr>
<td>Fan, Xiao</td>
</tr>
<tr>
<td>Fan, Jungwei</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>
Garcia, Jean  
Gartner, Daniel  
Gartrell, Kyungsook  
Garvin, Jennifer  
Garza, Maryam  
Gatwood, Justin  
Gaudioso, Carmelo  
Gazi, Asim  
Gaziel-Yablowitz, Michal  
Ge, Wendong  
Gehlot, Vijay  
Geifman, Nophar  
Geller, James  
Geng, Yimin  
Georgsson, Mattias  
Ghaffari, Meysam  
Ghalwash, Mohamed  
Gharibi, Gharib  
Ghosh, Debashis  
Ghosh, Debopriya  
Giannaris, Pericles  
Gibbings, Rima  
Gibbs, David  
Gibson, Richard  
Gibson, Bryan  
Ginige, Jeewani Anupama  
Giovannini, Andrea  
Girardi, Ivan  
Given, Barbara  
Glass, Kimberly  
Gligorijevic, Djordje  
Gobbel, Glenn  
Goetz, Kerry  
Gold, Rachel  
Gold, Jeffrey  
Goldstein, Ayelet  
Goldstein, Benjamin  
Goncalves Rezende Macieira, Tamara  
Gong, Yang  
Gong, Jen  
Goodwin, Rebecca  
Gordon, Geoffrey  
Gordon, William  
Goswami, Mononito  
Gottlieb, Assaf  
Goyal, Nikhil  
Graefe, Ronald  
Graetz, Ilana  
Grasso, Michael  
Greene, Sarah  
Greer, Melody  
Gregurick, Susan  
Griffin, Ashley  
Gross, Brian  
Grouin, Cyril  
Gruss, Calvin  
Gu, Yulong  
Gu, Yang  
Guidry, Alicia  
Guimarães, Marcelo  
Gundelach, Justin  
Guo, Yawen  
Guo, Xiaojie  
Gupta, Aditi  
Gupta, Samir  
Gupta, Akash  
Gupta, Deepak Kumar  
Gupta, Tanuj  
Gurol, Neslihan  
Gururaj, Anupama  
Guzman, Jenice  
Haberle, Tyler
<table>
<thead>
<tr>
<th>Name</th>
<th>First Name</th>
<th>Last Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hack, Lori</td>
<td>He, Lu</td>
<td>Hirko, Kelly</td>
</tr>
<tr>
<td>Hahn, Udo</td>
<td>He, Tiancheng</td>
<td>Hirschman, Lynette</td>
</tr>
<tr>
<td>Hajjar, Sarah</td>
<td>He, Yongqun</td>
<td>Hirst, Gillian</td>
</tr>
<tr>
<td>Haldar, Shefali</td>
<td>He, Huan</td>
<td>Ho, Joyce</td>
</tr>
<tr>
<td>Hall, Eric</td>
<td>He, Zhe</td>
<td>Ho, King Chung</td>
</tr>
<tr>
<td>Hamid, Zeyana</td>
<td>Heale, Bret</td>
<td>Hochheiser, Harry</td>
</tr>
<tr>
<td>Hamlish, Tamara</td>
<td>Heard, Kevin</td>
<td>Hoelscher, Stephanie</td>
</tr>
<tr>
<td>Han, Yan</td>
<td>Heermann Langford, Laura</td>
<td>Hoffman, Mark</td>
</tr>
<tr>
<td>Han, Sifei</td>
<td>Heider, Paul</td>
<td>Hoffman, James</td>
</tr>
<tr>
<td>Hanauer, David</td>
<td>Heisey-Grove, Dawn</td>
<td>Hoffman, Jeffrey</td>
</tr>
<tr>
<td>Harismendy, Olivier</td>
<td>Heitkemper, Elizabeth</td>
<td>Hogan, William</td>
</tr>
<tr>
<td>Harper, Ellen</td>
<td>Hellems, Martha</td>
<td>Hogarth, Mike</td>
</tr>
<tr>
<td>Harris, Kimberly</td>
<td>Heller, Oscar</td>
<td>Holder, Andre</td>
</tr>
<tr>
<td>Harry, Melissa</td>
<td>Hendrix, Nathaniel</td>
<td>Holl, Felix</td>
</tr>
<tr>
<td>Hart, Chip</td>
<td>Henry, Katharine</td>
<td>Hollberg, Julie</td>
</tr>
<tr>
<td>Hartzler, Andrea</td>
<td>Hensley Alford, Sharon</td>
<td>Holmgren, A Jay</td>
</tr>
<tr>
<td>Hasan, Md Mehedi</td>
<td>Herman, Daniel</td>
<td>Holsopple, Megan</td>
</tr>
<tr>
<td>Hassanzadeh, Hamed</td>
<td>Hettinger, Aaron</td>
<td>Homco, Juell</td>
</tr>
<tr>
<td>Hastings, Janna</td>
<td>Hicks, Kevin</td>
<td>Hong, Matthew</td>
</tr>
<tr>
<td>Hatef, Elham</td>
<td>Hide Prof., Winston</td>
<td>Hong, Julian</td>
</tr>
<tr>
<td>Hauser III, Ronald</td>
<td>Higbea, Raymond</td>
<td>Hoopes, Megan</td>
</tr>
<tr>
<td>Hayes, Corey</td>
<td>Hildebrandt, Gerhard</td>
<td>Horsky, Jan</td>
</tr>
<tr>
<td>Hays, David</td>
<td>Hill, Dave</td>
<td>Horvat, Christopher</td>
</tr>
<tr>
<td>He, Daqing</td>
<td>Himes, Blanca</td>
<td>Hosny, Abdelrahman</td>
</tr>
<tr>
<td>Hoyt, Robert</td>
<td>Iribarren, Sarah</td>
<td>Johnson, Marcus</td>
</tr>
<tr>
<td>-----------------------</td>
<td>------------------</td>
<td>--------------------------</td>
</tr>
<tr>
<td>Hribar, Michelle</td>
<td>Itoh, Sakiko</td>
<td>Johnson, Andrew</td>
</tr>
<tr>
<td>Hron, Jonathan</td>
<td>Ivory, Catherine</td>
<td>Johnson, Carl</td>
</tr>
<tr>
<td>Hsiao, Chun-Ju</td>
<td>Ize-Ludlow, Diego</td>
<td>Johnson, Elizabeth</td>
</tr>
<tr>
<td>Hsiao, Allen</td>
<td>Jackman, Kevon-Mark</td>
<td></td>
</tr>
<tr>
<td>hsieh, kang-lin</td>
<td>Jacobs, Jason</td>
<td>Jones, Barrett</td>
</tr>
<tr>
<td>Hsueh, Pei-Yun Sabrina</td>
<td>Jagannath, Swathi</td>
<td>Jones, JB</td>
</tr>
<tr>
<td>Hu, Stephanie</td>
<td>Jain, Nilesh</td>
<td>Jones, Josette</td>
</tr>
<tr>
<td>Hu, Zicheng</td>
<td>Jain, Nishant</td>
<td>Joo, Hyeon</td>
</tr>
<tr>
<td>Hu, Pengwei</td>
<td>Jamieson, Trevor</td>
<td>Joopudi, Venkata</td>
</tr>
<tr>
<td>Hu, Lu</td>
<td>Jang, Hyeju</td>
<td>Joshi, Shreekanth</td>
</tr>
<tr>
<td>Huang, Catherine</td>
<td>Jankovic, Ivana</td>
<td>Judson, Timothy</td>
</tr>
<tr>
<td>Huang, Ming</td>
<td>Jaynes, Heather</td>
<td>Jumbo, Adiebonye Eunice</td>
</tr>
<tr>
<td>Huang, Yan</td>
<td>Jeffery, Alvin</td>
<td>Jungbauer Jr, Walter</td>
</tr>
<tr>
<td>Hulse, Nathan</td>
<td>Jenders, Robert Allen</td>
<td></td>
</tr>
<tr>
<td>Hume, Samuel</td>
<td>Jenkins, Melinda</td>
<td>Kaelber, David</td>
</tr>
<tr>
<td>Hung, Veronica</td>
<td>Jeong, Eugene</td>
<td>Kakarmath, Sujay</td>
</tr>
<tr>
<td>Hussain, Mustafa Ibraheem</td>
<td>Jiang, Jie</td>
<td>Kalenderian, Elsbeth</td>
</tr>
<tr>
<td>Hyun, Sookyung</td>
<td>Jiang, Silis</td>
<td>Kalvesmaki, Andrea</td>
</tr>
<tr>
<td>Icardi, Michael</td>
<td>Jing, Xia</td>
<td>Kalyanaraman, Avinash</td>
</tr>
<tr>
<td>Ide, Tsuyoshi</td>
<td>Joffe, Erel</td>
<td>Kamala Raghavan, Sajeesh Kumar</td>
</tr>
<tr>
<td>Idnay, Betina</td>
<td>Johnson, Adriana</td>
<td>Kamdar, Maulik Rajendra</td>
</tr>
<tr>
<td>Introne, Josh</td>
<td>Johnson, Andrew</td>
<td>Kan, Mengyuan</td>
</tr>
<tr>
<td>Iott, Bradley</td>
<td>Johnson, Travis</td>
<td>Kandaswamy, Swaminathan</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Kang, Youjeong</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Kanga, Samuel Gichohi</td>
</tr>
<tr>
<td>Name</td>
<td>Name</td>
<td>Name</td>
</tr>
<tr>
<td>--------------------------------</td>
<td>---------------------------</td>
<td>---------------------------</td>
</tr>
<tr>
<td>Kovalchuk, Sergey</td>
<td>Lai, Jiaying</td>
<td>Leonardi, Giorgio</td>
</tr>
<tr>
<td>Krause, Kevin</td>
<td>Lai, Kenneth</td>
<td>Leroy, Gondy</td>
</tr>
<tr>
<td>Krause, Tim</td>
<td>Landman, Joshua</td>
<td>Lesh, Kathy</td>
</tr>
<tr>
<td>Krichevsky, Spencer</td>
<td>Landsman, David</td>
<td>Lesselroth, Blake</td>
</tr>
<tr>
<td>Krishnamurthy, Ashok</td>
<td>Lane, Steven</td>
<td>Leung, Tiffany</td>
</tr>
<tr>
<td>Krishnaswamy, Pavitra</td>
<td>Larson, Nicholas</td>
<td>Levandowski, Brooke</td>
</tr>
<tr>
<td>Krive, Jacob</td>
<td>Lau, Annie</td>
<td>Levy, Bruce</td>
</tr>
<tr>
<td>Kucharska-Newton, Anna</td>
<td>Lauria, Tara</td>
<td>Lewis, Deborah</td>
</tr>
<tr>
<td>Kuelbs, Cynthia</td>
<td>Laurio, Angela</td>
<td>Lewis, Paul</td>
</tr>
<tr>
<td>Kueper, Jacqueline</td>
<td>Leafe, Morgan</td>
<td>Li, Zheqi</td>
</tr>
<tr>
<td>Kuhn, Thomson</td>
<td>Leavy, Michelle</td>
<td>Li, Xin</td>
</tr>
<tr>
<td>Kulshrestha, Sujay</td>
<td>Lee, Kye Hwa</td>
<td>Li, Yan</td>
</tr>
<tr>
<td>Kumar, Manish</td>
<td>Lee, Dae Hyun</td>
<td>Li, Jiayun</td>
</tr>
<tr>
<td>Kumar, Karan</td>
<td>Lee, Adam</td>
<td>Li, Cheng</td>
</tr>
<tr>
<td>Kuo, Tsung-Ting</td>
<td>Lee, Jaehoon</td>
<td>Li, Jianfu</td>
</tr>
<tr>
<td>Kurc, Tahsin</td>
<td>Lee, Sanghoon</td>
<td>Li, Zhiguo</td>
</tr>
<tr>
<td>Kury, Fabricio</td>
<td>Lee, Scott</td>
<td>Li, Zhengyi</td>
</tr>
<tr>
<td>Kuttler, Kathryn Gibb</td>
<td>Lee, Young Ji</td>
<td>Li, Ruowang</td>
</tr>
<tr>
<td>Kuziemsky, Craig</td>
<td>Lee, Brian</td>
<td>Li, Fuhai</td>
</tr>
<tr>
<td>Kwatra, Japneet</td>
<td>Lee, Eva</td>
<td>Li, Ying</td>
</tr>
<tr>
<td>Kwon, Bum Chul</td>
<td>Lehmann, Harold</td>
<td>Li, Xiaojin</td>
</tr>
<tr>
<td>La Cava, William</td>
<td>Lehmann, Christoph</td>
<td>Liang, Lifan</td>
</tr>
<tr>
<td>Labkoff, Steven</td>
<td>Lei, Victor</td>
<td>Liang, Gongbo</td>
</tr>
<tr>
<td>Lacroix, Paulette</td>
<td>Lemke, Klaus</td>
<td>Liao, Frank</td>
</tr>
<tr>
<td>Name</td>
<td>Name</td>
<td>Name</td>
</tr>
<tr>
<td>--------------------------------</td>
<td>--------------------------------</td>
<td>--------------------------------</td>
</tr>
<tr>
<td>Liaw, Siaw-Teng</td>
<td>Liu, Xinran</td>
<td>MacKelfresh, Andy</td>
</tr>
<tr>
<td>Lichtner, Valentina</td>
<td>Lo, Ying-Chih</td>
<td>Macumber, Carol</td>
</tr>
<tr>
<td>Lillehaug, Svein-Ivar</td>
<td>Lomotan, Edwin</td>
<td>Madani, Sina</td>
</tr>
<tr>
<td>Lim Choi Keung, Sarah</td>
<td>Lopez-Campos, Guillermo</td>
<td>Madera, Martin</td>
</tr>
<tr>
<td>Lin, Anthony</td>
<td>Lorenzi, Virginia</td>
<td>Mahajan, Diwakar</td>
</tr>
<tr>
<td>Lin, Jia-Ling</td>
<td>Lou, Sunny</td>
<td>Mahajan, Satish</td>
</tr>
<tr>
<td>Lin, Zhen</td>
<td>Lourie, Eli</td>
<td>Maharathi, Biswajit</td>
</tr>
<tr>
<td>Lin, Ko-Hong</td>
<td>Lubin, Ira</td>
<td>Mahnke, Andrea Nicole</td>
</tr>
<tr>
<td>Ling, Albee</td>
<td>Luo, Xiao</td>
<td>Mahou, Malika</td>
</tr>
<tr>
<td>Liu, Xiong</td>
<td>Luo, Chongliang</td>
<td>Major, Vincent</td>
</tr>
<tr>
<td>Liu, Xu</td>
<td>Luo, Brooke</td>
<td>Malec, Scott</td>
</tr>
<tr>
<td>Liu, Sijia</td>
<td>Luyckx, Kim</td>
<td>Malik, Maria</td>
</tr>
<tr>
<td>Liu, Alex</td>
<td>Lyalin, David</td>
<td>Mallow, Jennifer</td>
</tr>
<tr>
<td>Liu, Kefei</td>
<td>Lybarger, Kevin</td>
<td>Maloney, Christopher</td>
</tr>
<tr>
<td>Liu, Guanghui</td>
<td>Lyle, Jay</td>
<td>Man, Jianing</td>
</tr>
<tr>
<td>Liu, Feifan</td>
<td>Lyles, Courtney</td>
<td>Manataki, Areti</td>
</tr>
<tr>
<td>Liu, Guodong</td>
<td>Lytle, Kay</td>
<td>Manders, Eric-Jan</td>
</tr>
<tr>
<td>Liu, Yongtai</td>
<td>Lyu, Tianchu</td>
<td>Manilich, Elena</td>
</tr>
<tr>
<td>Liu, Danlu</td>
<td>Ma, Andy Jinhua</td>
<td>Manion, Frank</td>
</tr>
<tr>
<td>Liu, Wanli</td>
<td>Ma, Meng</td>
<td>Manning, Matthew</td>
</tr>
<tr>
<td>Liu, Songzi</td>
<td>Ma, Xiaomeng</td>
<td>Manos, Eva LaVerne</td>
</tr>
<tr>
<td>Liu, Siru</td>
<td>Ma, Fenglong</td>
<td>Mao, Chengsheng</td>
</tr>
<tr>
<td>Liu, Hao</td>
<td>Mac Aonghuisa, Pol</td>
<td>Marafino, Ben</td>
</tr>
<tr>
<td>Liu, Mei</td>
<td>Macedo, Maysa</td>
<td>Marasinou, Chrysostomos</td>
</tr>
</tbody>
</table>
Marceglia, Sara
Marcial, Laura Haak
Marco-Ruiz, Luis
Marney, Heather
Marshall, Kyle
Marshall, Robert
Marszalek, Stephanie
Martens, Andy
Martin, Blake
Martin, Jacob
Martinez, William
Martini, Sharyl
Martins, Susana
Masese, Rita Vanessa
Masino, Aaron
Mathias, Patrick
Matin, Roni
Mavroudeas, Georgios
May, Sarah
Mazza, Kathleen
McClay, James
McClure, Robert
McCormick-Ricket, Iben
McCune, Jeannine
McGilchrist, Mark
McInnes, Bridget
McKillop, Mollie Marian
McKinley, Danette Waller
McPeek Hinz, Eugenia
Medina, Angeli
Meng, Frank
Menon, Priya
Metke Jimenez, Alejandro
Meyer, Anne-Marie
Miao, Zhuqi
Michalopoulos, George
Miksch, Tim
Milgrom, Zheng
Miller, Holly
Miller, Christopher
Miller, Timothy
Miller, Andrew
Min, Aehong
Mishra, Simita
Mishra, Shweta
Mishra, Meenakshi
Mishra, Rashmi
Mitchell, Jason
Mitchell, Elliot Griffith
Mittu, Ranjeev
Mobasher, Azadeh
Moeller, F. Gerard
Moen, Anne
Mohamed Ariff, Aminah Mazyin
Mohan, Santosh
Moldwin, Asher
Moldwin, Richard
Mombini, Haadi
Monahan, Rachel
Monsen, Karen
Moore, Carlton
Moores Todd, Tamara
Morassi Sasso, Ariane
Morid, Mohammad Amin
Morrow, Jon
Mosa, Abu Saleh Mohammad
Moshiri, Niema
Moss, Laura
Motiwala, Tasneem
Mottalib, Md Mozaharul
Mougin, Fleur
Movahedi, Faezeh
<table>
<thead>
<tr>
<th>Name</th>
<th>Name</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mowery, Danielle</td>
<td>Natsiavas Mr, Pantelis</td>
<td>Nong, Paige</td>
</tr>
<tr>
<td>Moye Jr., Jack</td>
<td>Naveed PhD, Hammad</td>
<td>Norquist, Craig</td>
</tr>
<tr>
<td>Mrosak, Justine</td>
<td>Nayebi, Amin</td>
<td>Norris, Amy</td>
</tr>
<tr>
<td>Mujib, Munif Ishad</td>
<td>Nestor, Jordan Gabriela</td>
<td>Novacek, Vit</td>
</tr>
<tr>
<td>Mukherjee, Sukrit</td>
<td>Neumann, Larissa</td>
<td>Novak, Laurie Lovett</td>
</tr>
<tr>
<td>Müller, Stephanie</td>
<td>Neumann, Victoria</td>
<td>O'Connor, Patrick</td>
</tr>
<tr>
<td>Mundt, Barbie</td>
<td>Neumuth, Thomas</td>
<td>Oeser, Alexander</td>
</tr>
<tr>
<td>Murcko, Anita</td>
<td>Neveol, Aurelie</td>
<td>Ogunyemi, Omolola</td>
</tr>
<tr>
<td>Murfin, Melissa</td>
<td>Newbold, Susan</td>
<td>Oh, Wonsuk</td>
</tr>
<tr>
<td>Murphy, Robert</td>
<td>Newman, Barry</td>
<td>O’Horo, John Charles</td>
</tr>
<tr>
<td>Mustata Wilson, Gabriela</td>
<td>Newman, David I</td>
<td>Ojo, Adebowale</td>
</tr>
<tr>
<td>Muth, Andrew</td>
<td>Newman-Griffis, Denis</td>
<td>Olatosoi, Bankole</td>
</tr>
<tr>
<td>Muthu, Naveen</td>
<td>Ng, Kenney</td>
<td>Oliwa, Tomasz</td>
</tr>
<tr>
<td>Myers, Risa</td>
<td>Ngo, Hoa</td>
<td>Omitaomu, Olufemi</td>
</tr>
<tr>
<td>Nadkarni, Girish Nitin</td>
<td>Nguyen, Thanh Minh</td>
<td>Omoloja, Abiodun</td>
</tr>
<tr>
<td>Nagaraj, Divya</td>
<td>Nguyen, Thanh</td>
<td>Orobaton, Nosa</td>
</tr>
<tr>
<td>Nagaraj, Gayathri</td>
<td>Nguyen, Minh</td>
<td>Orwoll, Benjamin</td>
</tr>
<tr>
<td>Nahm, Eun-Shim</td>
<td>Nguyen, Dac-Trung</td>
<td>Oskotsky, Tomiko</td>
</tr>
<tr>
<td>Nair, Rajeshwari</td>
<td>Nguyen, Vickie</td>
<td>Ostasiewski, Brian</td>
</tr>
<tr>
<td>Nakayama, Masaharu</td>
<td>Ni, Yizhao</td>
<td>Osterweil, Leon</td>
</tr>
<tr>
<td>Nakikj, Drashko</td>
<td>Nielson, Jeffrey</td>
<td>Ostovari, Mina</td>
</tr>
<tr>
<td>Nambudiri, Vinod</td>
<td>Ning, Xia</td>
<td>Ozaydin, Bunyamin</td>
</tr>
<tr>
<td>Nampally, Arun</td>
<td>Nocera, Luciano</td>
<td>Ozeran, Larry</td>
</tr>
<tr>
<td>Natarajan, Annamalai</td>
<td>Nolen, Kim</td>
<td>Ozery-Flato, Michal</td>
</tr>
</tbody>
</table>
Qi, Miao
Qiao, Zhi
Qu, Zhenhong
Quiroz, Juan Carlos
Rafalski, Matt
Ragaban, Nouran
Raghu, Vineet
Rahimian, Maryam
Rahman, Protiva
Raines, Greshundria
Raj, Shriti
Raja, Naveen
Rajabiyyazdi, Fateme
Rajamani, Sripriya
Raje, Satyajeet
Ramaswamy, Priya
Ramljak, Dusan
Ramos, S. Raquel
Rao, Arvind
Rashidee, Ali
Rasmussen, Luke
Rasmussen, Kelli
Rasooly, Irit
Rauscher, Richard
Rawat, Bhanu Pratap Singh
Reading Turchioe, Meghan
Recsky, Chantelle
Reddy, Madhu
Redley, Bernice
Reed, Robyn
Reeder, Blaine
Rehm, Gregory
Rehman, Mohamed
Rehman, Shakaib
Reid, Nick
Reimer, Andrew
Ren, Ziyu
Reyes Nieva, Harry
Reynolds, Tera
Reza, Faisal
Rhodes, Bryn
Riaz, Irbaz Bin
Ricarte, Ivan Luiz Marques
Richardson, Joshua E
Rienmüller, Theresa
Rios, Anthony
Rivard, Mary Jane
Rivera, Rebecca Lynn-Anne
Rizvi, Syed
Roberts, Pamela
Roderer, Nancy
Rodriguez, Victor
Rodriguez, Laritza
Rodriguez, Jorge
Rodríguez González, David
Rogers, Courtney
Rogers, James
Roller, Roland
Roman, Adrienne
Romano, Joseph
Ronquillo, Jay
Rosati, Robert
Rose, Rebecca
Rose, Christian
Rosenbloom, S. Trent
Rosenthal, Alex
Rosen-Zvi, Michal
Ross, Mindy
Ross, Stephen
Ross, Edgar
<table>
<thead>
<tr>
<th>Name</th>
<th>Name</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stella, Peter Alexander</td>
<td>Sward, Katherine</td>
<td>Taroni, Jaclyn</td>
</tr>
<tr>
<td>Stoffel, Michelle</td>
<td>Swoboda, Walter</td>
<td>Tavakoli, Abbas</td>
</tr>
<tr>
<td>Stojanovic, Konstantin</td>
<td>Syed, Sahr</td>
<td>Tavakoli Hosseinabadi, Maryam</td>
</tr>
<tr>
<td>Stonbraker, Samantha</td>
<td>Syeda-Mahmood, Tanveer</td>
<td>Tedla, Yacob</td>
</tr>
<tr>
<td>Stone III, Norman</td>
<td>Sylla, Issa</td>
<td>Temnikova, Irina</td>
</tr>
<tr>
<td>Strasberg, Howard</td>
<td>Sylvestre, Emmanuelle</td>
<td>Ter Meer, Lous</td>
</tr>
<tr>
<td>Strudwick, Gillian</td>
<td>Syrowatka, Ania</td>
<td>Thaker, Khushboo</td>
</tr>
<tr>
<td>Suarez Saiz, Fernando</td>
<td>Ta, Casey N</td>
<td>Thomas, Jason</td>
</tr>
<tr>
<td>Subbian, Vignesh</td>
<td>Tabernik, Ty</td>
<td>Thyvalikakath, Thankam</td>
</tr>
<tr>
<td>Subramaniam, Saranjah</td>
<td>Taft, Teresa</td>
<td>Tiase, Victoria</td>
</tr>
<tr>
<td>Sudhir Pillai, Parvathy</td>
<td>Tajgardoon, Mohammadamin</td>
<td>Tierney, Meghan</td>
</tr>
<tr>
<td>Suhaimi, Nurul</td>
<td>Talbert, Douglas</td>
<td>Times, Valéria Cesário</td>
</tr>
<tr>
<td>Suharwardy, Sanaa Hasan</td>
<td>Talmon, Geoffrey</td>
<td>Timkovich, Nick</td>
</tr>
<tr>
<td>Sujansky, Walter</td>
<td>Tan, Qingxiong</td>
<td>Timperi, Ralph</td>
</tr>
<tr>
<td>Sulieman, Lina</td>
<td>Tan, Hung-Jui</td>
<td>Toman, Rachelle</td>
</tr>
<tr>
<td>Sulley, Saanie</td>
<td>Tan, Xing</td>
<td>Tootooni, Samie</td>
</tr>
<tr>
<td>Sun, Tony</td>
<td>Tandon, Sanjeev</td>
<td>Topaloglu, Umit</td>
</tr>
<tr>
<td>Sun, Yingcheng</td>
<td>Tang, Ziyang</td>
<td>Topaz, Maxim</td>
</tr>
<tr>
<td>Sun, Carolyn</td>
<td>Tang, Paul</td>
<td>Torii, Manabu</td>
</tr>
<tr>
<td>Sun, Zhaonan</td>
<td>Tannier, Xavier</td>
<td>Tormey, Destinee</td>
</tr>
<tr>
<td>Suresh, Srinivasan</td>
<td>Tanon, Affaud Anais</td>
<td>Totzke, Michael</td>
</tr>
<tr>
<td>Surian, Didi</td>
<td>Tao, Donghua</td>
<td>Tourani, Roshan</td>
</tr>
<tr>
<td>Suryanarayanan, Parthasarathy</td>
<td>Tao, Carson</td>
<td>Toyoda, Shuichi</td>
</tr>
<tr>
<td>Sutariya, Bharat</td>
<td>Tao, Liang</td>
<td>Tran, Duong Trung</td>
</tr>
</tbody>
</table>
Trapp-Petty, Melissa  VanSchaik, Jack  Walton, Nephi
Trepp, Richard  VanZandbergen, Christine  Wan, Zhiyu
Trinidad, Marie  Vatani, Haleh  Wanderer, Jonathan
Tripp, Jacob  Vawdrey, David  Wang, Patrick
Tseng, Yi-Ju  Velez, Olivia  Wang, Ellen
Tsiknakis, Manolis  Veltri, Pierangelo  Wang, Ping
Tsou, Ching-Huei  Vera-Licona, Paola  Wang, Chen
Tu, Samson  Vert, Ann Leslie  Wang, Xun
Tumkur, Kashyap  Vincent, Adam  Wang, Ying
Tunnell IV, Harry  Visweswaran, Shyam  Wang, Haozhu
Turchin, Alexander  Vitali, Francesca  Wang, Ying
Turner, Kea  Volpe, Salvatore  Wang, Yanshan
Tweedey, Carolyn  Wagholikar, Kavishwar  Wang, Lu
Tyuryumina, Ella  Wahi, Monika  Wang, Xiaoyan
Ulapane, Nalika  Waitman, Lemuel  Wang, Amy
Umberfield, Elizabeth  Walcott, Aisha  Wang, Lucy Lu
Uptegraft, Colby  Walden, Anita  Wang, Liqin
Uzun, Ece  Walker, Lorne  Wang, Qinyong
Van Auker, Kimberly  Walker, Daniel  Wanyee, Steven
van Buchem, Marieke  Walker, James  Warnekar, Pradnya
Van Cain, Melissa  Walls, Ramona  Warner, Jeremy
van der Vegte, Anton  Walsh, Karen  Warzel, Denise
Van Schalkwyk, Johan  Walsh, Colin  Watkins, Michael
Vance, Mark  Walters-Threat, Lois  Weber, Jens
<table>
<thead>
<tr>
<th>Name</th>
<th>First Name</th>
<th>Last Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wei, Duo (Helen)</td>
<td>Wilson</td>
<td>Andrew</td>
</tr>
<tr>
<td>Wei, Wei</td>
<td>Winden</td>
<td>Tamara</td>
</tr>
<tr>
<td>Weiner, Stephanie</td>
<td>Windle</td>
<td>John</td>
</tr>
<tr>
<td>Weiner, Mark</td>
<td>Woeltje</td>
<td>Keith</td>
</tr>
<tr>
<td>Weinreb, Joanne</td>
<td>Wojtusiak</td>
<td>Janusz</td>
</tr>
<tr>
<td>Weir, Charlene</td>
<td>Wong</td>
<td>Anthony</td>
</tr>
<tr>
<td>Weiss, Doria</td>
<td>Woo</td>
<td>Kyungmi</td>
</tr>
<tr>
<td>Weiss, Jeremy</td>
<td>Wood</td>
<td>Elizabeth</td>
</tr>
<tr>
<td>Wen, Andrew</td>
<td>Wood</td>
<td>Grant</td>
</tr>
<tr>
<td>West, Vivian</td>
<td>Workman</td>
<td>Terri</td>
</tr>
<tr>
<td>West, Matthew</td>
<td>Wrendina</td>
<td>Mark</td>
</tr>
<tr>
<td>Whetzel, Patricia</td>
<td>Wright</td>
<td>Adam</td>
</tr>
<tr>
<td>Whitley, Eric</td>
<td>Wright</td>
<td>Aileen</td>
</tr>
<tr>
<td>Whitman, Greg</td>
<td>Wu</td>
<td>YiFan</td>
</tr>
<tr>
<td>Whittenburg, Luann</td>
<td>Wu</td>
<td>Honghan</td>
</tr>
<tr>
<td>Wickramasinghe, Nilmini</td>
<td>Wu</td>
<td>Yonghui</td>
</tr>
<tr>
<td>Wiley, Ken</td>
<td>Wu</td>
<td>Shinyi</td>
</tr>
<tr>
<td>Wilk, Szymon</td>
<td>Wu</td>
<td>Huanmei</td>
</tr>
<tr>
<td>Wilkinson, Katie</td>
<td>Wu</td>
<td>Danny TY</td>
</tr>
<tr>
<td>Wilkinson, Thomas</td>
<td>Wu</td>
<td>Joy Tzung-yu</td>
</tr>
<tr>
<td>Williams, Nick</td>
<td>Xi</td>
<td>Wenna</td>
</tr>
<tr>
<td>Williams, Marc</td>
<td>Xie</td>
<td>Feng</td>
</tr>
<tr>
<td>Williams, Karmen</td>
<td>Xie</td>
<td>Sherrie</td>
</tr>
<tr>
<td>Wilson, Kala</td>
<td>Yahagi</td>
<td>Naohisa</td>
</tr>
<tr>
<td>Yakubu, Abubakari</td>
<td>Yan</td>
<td>Xiaowei</td>
</tr>
<tr>
<td>Yan, Chao</td>
<td>Yan</td>
<td>Yao</td>
</tr>
<tr>
<td>Yang, Jie</td>
<td>Yang</td>
<td>Chengliang</td>
</tr>
<tr>
<td>Yang, Jianji</td>
<td>Yang</td>
<td>Jie</td>
</tr>
<tr>
<td>Yang, Tzu-I</td>
<td>Yang</td>
<td>Christopher</td>
</tr>
<tr>
<td>Yang, Rumei</td>
<td>Yao</td>
<td>Robert</td>
</tr>
<tr>
<td>Yao, Xinghua</td>
<td>Ye</td>
<td>Ping</td>
</tr>
<tr>
<td>Ye, Jiancheng</td>
<td>Ye</td>
<td>Cheng</td>
</tr>
<tr>
<td>Ye, Cheng</td>
<td>Yen</td>
<td>Po-Yin</td>
</tr>
<tr>
<td>Yim, Wen-wai</td>
<td>Yoon</td>
<td>Jinsung</td>
</tr>
<tr>
<td>Yousif, Zaid</td>
<td>Yu</td>
<td>Yue</td>
</tr>
<tr>
<td>Yu, Jingzhi</td>
<td>Yu</td>
<td>Zehao</td>
</tr>
<tr>
<td>Yu, Fei</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Name</td>
<td>Name</td>
<td>Name</td>
</tr>
<tr>
<td>---------------------------</td>
<td>---------------------------</td>
<td>---------------------------</td>
</tr>
<tr>
<td>Yu, Sean Chonghwan</td>
<td>Zhang, Lin</td>
<td>Zong, Nansu</td>
</tr>
<tr>
<td>Yu, Yue</td>
<td>Zhang, Huabing</td>
<td>Zong, Wei</td>
</tr>
<tr>
<td>Yu, Deahan</td>
<td>Zhang, Ping</td>
<td>Zorc, Joseph</td>
</tr>
<tr>
<td>Yuan, Zheng</td>
<td>Zhao, Qingqing</td>
<td>Zubriski, Mark</td>
</tr>
<tr>
<td>Yurk, Robin</td>
<td>Zhao, Qinpei</td>
<td></td>
</tr>
<tr>
<td>Yusuf, Rafeek</td>
<td>Zhao, Yiqing</td>
<td></td>
</tr>
<tr>
<td>Zang, Huaiyu</td>
<td>Zhao, Juan</td>
<td></td>
</tr>
<tr>
<td>Zayas-Cabán, Teresa</td>
<td>Zhao, Shilin</td>
<td></td>
</tr>
<tr>
<td>Zech, John</td>
<td>Zhen, Yi</td>
<td></td>
</tr>
<tr>
<td>Zelcer, John</td>
<td>Zheng, Ling</td>
<td></td>
</tr>
<tr>
<td>Zeng, Xiaoming</td>
<td>Zheng, Hanming</td>
<td></td>
</tr>
<tr>
<td>Zhang, Zuoyi</td>
<td>Zheng, Zhi</td>
<td></td>
</tr>
<tr>
<td>Zhang, Mengnan</td>
<td>Zheng, Fengbo</td>
<td></td>
</tr>
<tr>
<td>Zhang, Yu</td>
<td>Zheng, Jiaping</td>
<td></td>
</tr>
<tr>
<td>Zhang, Yuan</td>
<td>Zheng, Chunlei</td>
<td></td>
</tr>
<tr>
<td>Zhang, Yuji</td>
<td>Zhou, Shuxin</td>
<td></td>
</tr>
<tr>
<td>Zhang, Zhan</td>
<td>Zhou, Xiang</td>
<td></td>
</tr>
<tr>
<td>Zhang, Zilong</td>
<td>Zhou, Sicheng</td>
<td></td>
</tr>
<tr>
<td>Zhang, Yili</td>
<td>Zhou, Leming</td>
<td></td>
</tr>
<tr>
<td>Zhang, He</td>
<td>Zhu, Ping</td>
<td></td>
</tr>
<tr>
<td>Zhang, Jianqiu</td>
<td>Zhu, Weiguo</td>
<td></td>
</tr>
<tr>
<td>Zhang, Yi</td>
<td>Zhu Dr, Vivienne</td>
<td></td>
</tr>
<tr>
<td>Zhang, Wenhui</td>
<td>Zikos, Dimitrios</td>
<td></td>
</tr>
<tr>
<td>Zhang, Ziqi</td>
<td>Zimmerman, Lindsay</td>
<td></td>
</tr>
</tbody>
</table>
NOTICE

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

Julia Adler-Milstein, PhD¹, Andrew Olson, MD², Charlene Weir, PhD, MS³, Benjamin Rosner, MD, PhD¹, Robert El-Kareh, MD, MPH, MS⁴

¹University of California San Francisco, San Francisco, CA; ²University of Minnesota, Minneapolis, MN; ³University of Utah, Salt Lake City, UT; ⁴University of California San Diego, La Jolla, CA

Abstract

Timely and accurate diagnosis is a cornerstone of effective and efficient care. Mastery of clinical diagnosis requires ongoing learning and calibration of performance. Our current systems are not well-suited to providing the diagnostic outcome feedback to enable this calibration. The evolving functionality of EHRs has the potential to provide such feedback to practicing clinicians, which would be a significant step toward achieving the goal of life-long diagnostic improvement. Our panel brings together a diverse set of researchers focused on different aspects of improving the diagnostic process using health information technology. The panel will conclude with a conversation about next steps that can be taken in the short term to harness our EHR functionality to better support clinician diagnostic calibration. Learning objectives include: 1) Identify the types of information required for effective diagnostic calibration; 2) Describe the limitations and barriers of current EHRs to provide useful diagnostic outcome feedback; and 3) Describe initial efforts to improve the ability of current EHRs to support ongoing diagnostic calibration of trainees and practicing clinicians.

Panel Description

Diagnostic decision-making takes years to master and clinicians expect to improve their diagnostic skills through practice. However, the learning process requires that clinicians become aware of the outcomes of their diagnostic decisions and identify opportunities to improve in the future. Our healthcare system has become fragmented, with frequent patient handoffs and multiple specialties involved in care that limit the ability for clinicians to see diagnostic outcomes. This lack of outcome feedback is a critical barrier to diagnostic calibration, and thus physicians remain poorly calibrated and often overconfident.(1, 2) Physicians in training recognize the importance of outcome feedback, but encounter significant barriers to obtaining it reliably.(3) Feedback can take a variety of forms, with differing impacts on motivation and learning (both positive and negative). For example, normative feedback (comparison to peers) can be aversive, but also motivating. Whereas task feedback directs attention and increases involvement, it also involves more effort. The technological environment of clinical practice limits information and communication, thereby affecting both diagnostic decision-making and calibration. Healthcare organizations can take steps to enhance their EHR system to support diagnosis and fully utilize the power of health information technology. Practicing clinicians may also require specific training to take full advantage of this technology to improve their diagnostic processes over time.(4)

In this panel, we will describe aspects of how EHRs affect the diagnostic process and will provide practical suggestions for how EHR functionality can be used to facilitate life-long diagnostic calibration. Each speaker will present their work with a focus on how EHRs affect clinicians’ visibility into their own diagnostic processes and outcomes. The four presentations will be followed by a facilitated conversation to discuss the challenges and opportunities of...
developing EHR-based systems to support diagnostic calibration. The panelists represent different institutions (U of Minnesota, U of Utah, UCSF, UCSD) and bring diverse perspectives on the topic. Specifically:

**Julia Adler-Milstein, PhD** (moderator) will introduce the panelists and the topic. She will describe the issue of diagnostic calibration within the National Academy of Medicine model of the diagnostic process. She will present a framework describing current efforts to improve diagnostic calibration and how they are supported by EHRs and EHR data. She will then describe how each of the panelists’ work fits within that framework. Following the presentations, Dr. Adler-Milstein will facilitate the discussion between the audience and panel.

**Andrew Olson, MD** (panelist) will present the results from two projects related to diagnostic calibration. The first project, funded by the Gordon and Betty Moore Foundation, describes the lessons learned from discussing practices of feedback and calibration from other industries (e.g., aviation, meteorology and athletics). These discussions informed the creation of a diagnostic calibration framework. Dr. Olson will discuss how the EHR fits into that framework. The second project involves a non-EHR based educational project to provide diagnostic outcome feedback to medical trainees. Lessons learned from this project can help inform ways to use data within the EHR to create a scalable and sustainable diagnostic feedback system.

**Charlene Weir, PhD, MS** (panelist) will draw on her background in cognitive science and informatics and extensive experience applying cognitive theories to the design of health information technology to describe current barriers and opportunities to use EHRs to facilitate diagnostic calibration. She will integrate the psychological literature on feedback with EHR literature, focusing on how the lack of specification regarding diagnoses can impact “upstream” clinicians which in turn can have a significant impact on diagnostic performance. History and subjective data are critical to accurate diagnosis and pose particular informatics challenges—much of subjective data provided by patients and diagnostic assessments of clinicians are found in large sections of free text which are nearly impossible to find and hard to integrate into a bigger picture. Clinician workarounds for these challenges involve verbal communication, but much of this information is lost, especially during transitions of care. Despite these challenges, Dr. Weir will discuss several potential avenues to improve clinician-EHR collaboration in diagnosis. These include improved feedback systems, more effective data visualization and more helpful representation of clinical problems.

**Benjamin Rosner, MD, PhD** (panelist) will present results from work to use EHR metadata to improve patient attribution—assigning patient-level metrics to specific clinicians. Optimizing attribution is crucial to ensure that each clinician receives diagnostic feedback on the correct set of patients. Ideally, this attribution can be done in an automated manner based on who has taken actions in the patient chart (e.g., who wrote orders, who viewed the chart). In some settings, this will be more straightforward than others. For example, in the ambulatory care setting, a primary care provider (PCP) would likely own most or some degree of attribution, but this may be diagnosis dependent. Where specialists are involved, diagnostic accuracy may be a shared process, involving both the timeliness of referral by the PCP and referral to the correct specialty by the PCP (e.g., refer to Rheumatology vs Orthopedics for joint pain), as well as the specialist arriving at the correct diagnosis in a timely manner. In the inpatient setting, attribution based on who has interacted with the EHR for a given patient grows more complex. A hospitalist may be on service for a week and then hand off a patient’s care to another hospitalist, so both might be attributable to some degree. In team-based care, such as that associated with trainees (interns, residents), using the EHR to determine attribution is also complex. Dr. Rosner will discuss an approach to account for team-based care using EHR metadata from UCSF in which attribution for each patient on the team’s census can be quite accurately determined.

**Robert El-Kareh, MD, MPH, MS** (panelist and organizer) will describe initial results from an EHR dashboard that automates patient diagnostic outcome feedback to emergency medicine (EM) and internal medicine (IM) clinicians within the hospital. The presentation will illustrate how the design, implementation and educational issues raised by the preceding panelists were considered in an active, EHR-enabled feedback system within an academic medical center. Dr. El-Kareh will describe the evolution of the current system and how it is being used to support both targeted educational opportunities for trainees and ongoing reflection of patient outcomes by all clinicians within those specialties.
Topic Rationale

Diagnostic calibration is an essential aspect of high-quality and efficient care, but EHR design has not focused on supporting this process. Evolving functionality of EHRs and the data they contain can be harnessed to facilitate ongoing assessments of diagnostic outcomes at the clinician-level. However, the most effective methods for pursuing this work are yet to be determined. We anticipate that the audience for this panel will be clinical informaticists representing a wide range of healthcare settings who are working to support diagnostic excellence and its impact on patient safety, organizational efficiency and provider well-being.

Statement of Agreement to Participate

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

Panel Participant Details

Julia Adler-Milstein, PhD. Director, Center for Clinical Informatics & Improvement Research. Professor, Department of Medicine, University of California San Francisco
Andrew Olson, MD. Associate Professor of Medicine & Pediatrics, University of Minnesota Medical School
Charlene Weir, PhD, MS. Professor of Biomedical Informatics, University of Utah
Benjamin Rosner, MD, PhD. Associate Professor, Department of Medicine, University of California San Francisco
Robert El-Kareh, MD, MPH, MS. Associate Professor, Department of Medicine, University of California San Diego

References

Abstract

COVID-19 has stimulated innovation and collaboration on an unprecedented global scale. The COVID-19 Knowledge Accelerator (COKA) Initiative is an international volunteer organization with over a hundred participants and 11 active working groups advancing standards, systems, terminologies, metadata frameworks, tools, and demonstration projects for computable expression of biomedical knowledge. COKA is an evolution of the Fast Healthcare Interoperability Resources (FHIR®) for Evidence Based Medicine (EBM) Knowledge Assets (aka EBMonFHIR) project. These organizations are working to find solutions for the current inefficiencies in the scientific dissemination systems we use today. Using structured (computable) results data directly from research and research publications will greatly accelerate evidence synthesis by reducing redundant efforts and enhancing interoperability. A panel of COKA leaders with diverse clinical informatics experience and research endeavors will share their most recent accomplishments and challenges. Participants will learn what it takes to make science computable and how to participate in the COKA initiative.

Panel Description

Our panel brings diverse perspectives from clinical and research informatics, systems engineering, epidemiology, and evidence-based medicine. We will present an overview of the COKA Initiative and highlights of the most impactful developments to date and the most pressing challenges to solve. The panel will engage the audience in participation during the presentation and facilitate ongoing participation after the conference.

Brian S. Alper, MD, MSPH (the overall project lead) will serve as moderator and introduce the origins of the COKA Initiative. In addition to being project lead for COKA, Dr. Alper is the CEO of Computable Publishing LLC which is working to create a paradigm shift in the way we interact with scientific literature. Dr Alper will guide the panel through a discussion of functional applications to make science computable.

Harold Lehmann, MD, PhD is professor of informatics at Johns Hopkins Bloomberg School of Public Health. His research interests span medical informatics, evidence-based medicine, decision making, biostatistics, decision analysis and Bayesian communications. Dr. Lehmann will introduce issues encountered in efforts of the National COVID Cohort Collaborative (N3C) to systematically collect data from electronic health records and harmonize these data as a centralized resource for collaborative research.

Joanne Dehnbostel, MS, MPH is an epidemiologist and former laboratory scientist and is a project leader for Code System Development. She will introduce the challenges with lacking structured vocabularies for science, the open protocols for global collaborative development, and progress achieved in defining code systems for risk of bias, study design, statistic type, and statistical model concepts.

Andrey Soares, PhD, is an Assistant Professor in the Department of Medicine at the University of Colorado Anschutz Medical Campus. His research interests include data interoperability, clinical decision support systems, natural language processing and data analytics with the goal of developing applications and digital solutions to support evidence-based decision making. Dr. Soares will introduce some of the user interfaces and APIs used and developed by the COKA project, and share the current challenges in next-stage developments.

Vignesh Subbian, PhD is a Joint Assistant Professor of Biomedical Engineering and Systems and Industrial Engineering at the University of Arizona. Dr. Subbian also leads the Communications Work Group within the COVID-19 Knowledge Accelerator. Dr Subbian will show the most recent project developments (such as a systematic meta-review analyzing systematic reviews of steroid therapy for COVID-19 and converting all the data to computable evidence) and discuss the challenges in human-machine interaction and how we are overcoming those challenges.

The panel discussion will be a lively Q&A session based on the interests of the audience. In addition the moderator will facilitate a structured dialog that will follow a format approximating:

1) Audience participants are asked to report an unresolved challenge in communicating (finding, transforming, or sending) scientific evidence.
2) The group discusses what may be needed to overcome the challenge, such as a structural element in the schema for data exchange, a change in controlled vocabularies, or an improvement to the instructional materials explaining how to use a particular tool.

3) The group discusses alternatives for overcoming the challenge.

4) The group decides a suggested plan for next steps to solve the challenge or further investigate solutions.

**Importance and Focus**

The topic of COVID-19 is timely, urgent, needed, and attention grabbing. The topic of scientific knowledge acceleration is timely with recent advances in the Mobilizing Computable Biomedical Knowledge community, urgent as it applies immediately to COVID-19 knowledge, needed because it is not happening naturally as we continue to operate in a document-based world not using computers to their full potential for science communication, and attention grabbing because making science computable is a direct interest to anyone engaged in medical informatics.

The participants who would benefit most are researchers and knowledge management professionals who would benefit from being able to express biomedical evidence in computable (interoperable, re-usable, reproducible, shareable) formats.

**Discussion Questions to Enhance Audience Participation**

What would you be able to do if you could get scientific knowledge in a computable form?

What would you do if you could share scientific knowledge in a computable form?

Why are you not sharing scientific knowledge in a computable form now?

If you are, what makes it difficult?

How can you be part of the solution?

**Statement of Participation**

All participants have agreed to take part on the panel.

**References**


Using Specific COVID-19 Targets and Patient Care Settings as a Springboard for Driving Global Improvements in the Learning Health System Cycle for COVID-19 and Beyond

Steve Bernstein, BS1; Jerome A. Osheroff, MD2; Julia Skapik, MD3; Christopher Tignanelli, MD4; Matthew Burton, MD5

1Agency for Healthcare Research and Quality, Rockville, MD; 2TMIT Consulting, LLC, Naples, FL; 3National Association of Community Health Centers, Bethesda, MD; 4Department of Surgery, University of Minnesota, Minneapolis, MN; 5Apervita, Inc., Chicago, IL

Abstract

AHRQ formed the ACTS COVID-19 Evidence to Guidance to Action Collaborative in early 2020 as a learning community to help participants respond to the emerging pandemic. This Collaborative was an outgrowth of the AHRQ evidence-based Care Transformation Support (ACTS) initiative, which had spent the previous year developing a shared vision for health IT-enabled, evidence-informed care delivery and transformation, and a draft stakeholder-driven Roadmap for broadly achieving this vision. This panel will outline AHRQ’s goals and process for delivering broad value from the Collaborative and provide perspectives from three participants on how the Collaborative supported their efforts to improve care processes and outcomes for COVID-19 and beyond. After panelist presentations, the second session half will be interactive discussion with attendees (likely including other Collaborative participants) to surface opportunities to leverage efforts such as the Collaborative to accelerate care transformation within and across organizations.

Introduction

In the U.S. healthcare system, the flow of data, evidence, knowledge, and tools to support critical decisions and actions is highly siloed and fragmented. This leads to preventable problems with care and patient outcomes, unnecessary expenditures, suboptimal patient experience, and overburdened clinicians.1 While there have been major investments in health information technology (IT),2 value from these investments,3 and efforts to standardize decision support4,5 there remains a chasm between the degree of transformation achieved in other industries and that seen in healthcare. The COVID-19 pandemic has dramatically highlighted the urgent need to produce a virtuous learning health system (LHS) cycle where the flow from data to evidence to guidance to action and back to data continually improves. Ultimately, the goal – for COVID and beyond – is to achieve the Quadruple Aim6 to simultaneously enhance patient experience, improve population health, reduce costs, and improve clinician experience in care delivery.

The Agency for Healthcare Research and Quality (AHRQ) launched the AHRQ evidence-based Care Transformation Support (ACTS) Initiative7 in 2019 to develop a Roadmap for how to improve access to and use of resources that AHRQ and others provide to improve care delivery and transformation toward the Quadruple Aim. The central goal for ACTS is to ensure that evidence, knowledge, and guidance resources from AHRQ and others are optimally findable, accessible, interoperable, reusable (FAIR) for all who can benefit from them. To achieve this goal, ACTS established a Stakeholder Community to help outline the current state of evidence-informed care delivery and transformation, a shared desired future vision, and a proposed roadmap to get there.

In early 2020, AHRQ launched the ACTS COVID-19 Evidence to Guidance to Action Collaborative in response to the emerging pandemic as a learning community helping participants improve the LHS cycle for COVID-19 and beyond. This collaborative was formed as a step towards implementing the draft ACTS Roadmap approach to achieve the future vision. The central goals of the ACTS COVID-19 Evidence to Guidance to Action Collaborative are:

1) Cross-fertilize.accelerate efforts to develop & deliver COVID-19 evidence-based guidance and tools to patients and care teams
2) Measurably improve care and outcomes for selected targets and settings, and promote scaling to many others
3) Advance tools, standards, and collaborations needed for a knowledge ecosystem that supports LHSs (See Figure)
Leveraging the ACTS COVID Collaborative to Improve Care Delivery/Outcomes and the LHS Cycle

The Collaborative efforts are anchored by a diverse community of stakeholders who meet weekly as a learning community. Collaborative participants include care delivery organizations (CDOs) who are developing, implementing, and evaluating living CDS interventions focused on specific COVID-19 issues, health IT suppliers, standards developers, evidence synthesis and guidance developers, patient advocates, and representatives from many other stakeholder groups. Efforts to improve activities around the knowledge ecosystem/LHS cycle, and the results these activities produce, are centered on specific clinical COVID-19 patient care improvement targets (e.g., severity assessment/triage, use of anticoagulants, and managing Post-acute Sequelae of COVID-19 (“Long COVID”)) being addressed in specific care delivery organizations (e.g., Veterans Health Administration, University of Minnesota, OHSU, health centers).

Collaborative participants gather weekly by web meeting and use online collaboration tools and additional ad hoc meetings to foster learning and collaboration toward improving participants’ ecosystem-related efforts and results, and to define strategies for broader ecosystem cycle enhancements. Organizations focused on the specific topics share their strategies, challenges, successes, needs, insights, tools and approaches. Other participants provide feedback, support, and collaboration, and also benefit from valuable learning shared and related discussions. These activities are seeding execution of the proposed ACTS Roadmap by providing laying foundations for five core Roadmap activities:

- Seeding cross-stakeholder coordination efforts
- Defining and laying the foundation for enhanced ecosystem cycle infrastructure, standards, and tools
- Providing concept demonstrations for enhanced/computable evidence and guidance content & processes
- Providing concept demos for enhanced guidance implementation
- Laying the foundation, including best practices, for evaluation, planning, piloting, scaling

Proposed Panel

The panel’s goal is to build on the valuable dialogue, activities, and collaboration reflected in the ACTS COVID-19 Collaborative to benefit more organizations from this work and accelerate widespread progress toward the ACTS future vision. Annual Symposium attendees typically reflect many key LHS stakeholder groups, and this panel is aimed at such participants that have care transformation – for COVID-19 and beyond – as a focus.

Mr. Bernstein is the AHRQ project officer for ACTS and will describe AHRQ’s goals and context for ACTS and the COVID-19 Collaborative. Dr. Osheroff chairs the Collaborative and will describe the Collaborative strategy and results. Other panelists who are participants will discuss how they have benefitted from the Collaborative, including their activities, learnings and results related to the knowledge ecosystem, and implications from these for the audience:

Dr. Julia Skapik, National Association of Community Health Centers (NACHC), will discuss efforts in health centers related to improving care and outcomes for COVID-19 including Post-acute Sequelae for COVID-19 (PASC/‘Long COVID’), focusing on their challenges, needs and successes related to improving the knowledge
ecosystem cycle for this target. She will also address how the ACTS Collaborative has supported these efforts and implications for others. Dr. Christopher Tiganelli, University of Minnesota (UMN), will discuss efforts with the 12 hospitals in the UMN health system related to anticoagulation use and outcomes for hospitalized patients with COVID-19, also focusing on their challenges, needs and successes related to improving the knowledge ecosystem cycle for this target. He will also address how the ACTS Collaborative has supported these efforts and implications for others. Dr. Matthew Burton, Apervita, lead for the C19HCC Digital Guideline Working Group and contributor the HL7 FHIR Clinical Practice Guidelines (“CPG on FHIR”) Implementation Guide will discuss work to put this Implementation Guide into action with ACTS Collaborative participants (including UMN) to provide living, standards-based computable guidance for COVID-19 targets. He will also discuss challenges, needs, success and opportunities for leveraging initiatives such as the ACTS COVID Collaborative to make standards-based computable guidance widely available for many COVID-19 targets and beyond – including implications for others.

The panel’s second half will be devoted to interactive discussion among panelists and audience members (likely to include other Collaborative participants) to surface opportunities to further accelerate efforts to develop and deliver evidence-based guidance and tools for COVID-19 and beyond to patients and care teams, while simultaneously improving each of the other components of the LHS cycle. Prompts for this discussion include:

- Where do your efforts lie on the LHS cycle, and what challenges have you faced receiving inputs from preceding steps in the cycle, producing your deliverables, and optimizing their value in subsequent steps in the cycle?
- Are there opportunities to enhance the efficiency and effectiveness of your efforts through approaches, tools, standards, and collaborations as outlined in this panel?
- What would specific synergies between your efforts/plans and the Collaborative’s work look like, and what next steps to cultivate them would be helpful?

All participants have agreed to take part in the panel.

Conclusion

The COVID-19 pandemic has focused sharp attention on how far healthcare is from achieving LHSs and the urgency and high stakes for closing this gap. The ACTS COVID Collaborative is taking important steps in this direction, centering its efforts on specific care delivery organizations and their efforts related to addressing specific COVID-19 targets – in ways that can scale learning and other results to many other settings and targets. This interactive panel will build on this work and leave attendees with valuable insights and opportunities to advance their efforts in this area in ways that further accelerate progress toward a shared LHS future vision.

References

Representation Requires Intentionality: Our Journey to Creating a Diverse Informatics Workforce

Tiffani J. Bright, PhD, FACMI1, Chinyere Agunwa, PhD2, Kim Unertl, PhD, FACMI3, Oliver J. Bear Don’t Walk IV, MS4, Yalini Senathirajah, PhD5

1IBM Watson Health, Cambridge, MA; 2IBM Research, San Jose, CA; 3Vanderbilt University Medical Center, Nashville, TN; 4Columbia University, New York, NY; 5University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania

Abstract

Cultivating a diverse and inclusive workforce is critical for the biomedical informatics profession, both from a perspective of fairness and equity and to avoid technology that fails to address concerns of significant portions of the population or that introduces unintentional biases. However, having diversity and representation in the informatics workforce requires intentionality and deliberate action to combat systemic inequalities and requires focused efforts to ensure inclusive professional environments. This challenge necessitates active engagement in order to continue our work of transforming health and healthcare to ensure equitable outcomes. This panel reports on the experiences of informaticians and data scientists across academic and industry settings in creating diverse and inclusive environments. The learning objectives are to: 1) describe why a diverse workforce and inclusive working environments are critically important goals of the informatics field; 2) gain an understanding of lessons learned and best practices for broadening participation in informatics; and to 3) learn about resources to support efforts to broaden participation in informatics.

Introduction

Biomedical informatics presents many new opportunities and career directions. However, the field still lacks racial and ethnic diversity and representation; these disparities are present in doctoral programs1. When representation is missing from the informatics workforce, key voices are also missing in the design, implementation, and evaluation of technologies. The absence of these voices may lead to failures in scientific objectivity and the potential loss of innovative solutions to some of the most critical health and healthcare problems we face today, as well as increasing the potential for unintentional incorporation of biases into the design of technology-based interventions. Thus, building a diverse informatics workforce is necessary not only for the profession that will design solutions to solve health care problems, but also for transforming health and healthcare towards equity-centered care. These challenges are not unique to our field; it is time for informatics professionals, individually and collectively, to develop and implement strategies that will build and support diverse and inclusive environments5.

This panel will provide an overview of lessons learned during efforts to broaden the informatics workforce, based on speaking engagements at national conferences, the development of internship programs, and the creation of student outreach events and community college workforce programs. This panel should be of interest to informatics professionals who are interested in: 1) understanding practical strategies needed to build a diverse informatics workforce and 2) learning about resources to support those efforts. Each panelist will talk for 10-12 minutes and cover the following:

1. Why diversity in informatics matters to health and healthcare
2. Intersectionality and how it increases barriers and challenges related to exclusion
3. Pathways into informatics and institutional issues
4. The importance of mentoring, sponsorship, and career advancement

The moderator and panelists include:
Tiffani J. Bright, PhD, FACMI, Biomedical Informatician, IBM Watson Health

Dr. Bright is an informatician with expertise in the development and evaluation of health information technologies and clinical decision support. For the last 15 years, Dr. Bright has built a career spanning academia, public health,
government, and industry, and using her background and expertise as a platform to create a diverse and inclusive informatics and STEM workforce through local and national initiatives. She is also the founder of the AMIA First Look Program, which introduces undergraduate women to the informatics field. She serves as Chair of the AMIA Diversity, Equity, and Inclusion Task Force, the AMIA Board of the Directors, and the Women in AMIA Steering Committee. She is a Fellow of the American College of Medical Informatics (ACMI) and a blackcomputeHER Fellow.

Chinyere Agunwa, PhD, Research Staff Member, IBM

Dr. Agunwa is a Research Staff Member at IBM. She joined IBM Research in December 2015 as a medical image analysis and machine learning researcher on the Medical Sieve Radiology Grand Challenge project. She works on machine learning applications. Dr. Agunwa received her MS and PhD degrees in Electrical Engineering from Stanford University. She graduated with a double major in Computer and Electrical Engineering, and double minor in Computer Science and Mathematics from Louisiana State University (LSU). She has spoken at the Grace Hopper Celebration of Women in Computing, mentored high school girls through the Girls Who Code program, and led student outreach panels.

Kim Unertl, PhD, FACMI, Associate Professor of Biomedical Informatics, Vanderbilt University Medical Center

Dr. Unertl directs the Vanderbilt Biomedical Informatics Summer Program to engage high school and undergraduate students in biomedical informatics research. Her research focuses on clinical workflow, technology usability, and women’s health. She has mentored almost 20 high school and undergraduate students in summer research experiences over the last 10 years and has also mentored medical and graduate students. As part of her summer internship program work, Dr. Unertl developed and implemented a professional development program for summer interns based on the Entering Research curriculum from the Center for the Improvement of Mentored Experiences in Research. She is one of the co-founders of the AMIA High School Scholars Program and also serves on AMIA’s Education Committee.

Oliver J. Bear Don’t Walk IV, MS. PhD student in Biomedical Informatics, Columbia University

Oliver will moderate the panel and describe efforts to increase representation of persons excluded because of their ethnicity or race (PEERs) in STEM fields. Oliver is a PhD student studying the intersection of ethics, bias, and clinical natural language processing (NLP). Their work has been shaped by their Apsáalooke background and an emphasis on using education to give back to vulnerable populations. Oliver’s thesis will use NLP to recover social and behavioral determinants of health from clinical notes and provide best practices for ethically using such information to support healthcare and research. Oliver has served as a mentor and instructor for programs such as IndigiData and the Columbia DBMI Summer Research Program, aimed at increasing the representation of PEER students in STEM fields.

Yalini Senathirajah, PhD. Associate Professor in the Department of Biomedical Informatics, University of Pittsburgh School of Medicine.

Dr. Senathirajah has experience teaching and mentoring students and non-trainees seeking to enter the field, at both majority minority institutions (e.g., Downstate medical center, 85% minority), ivy league academic institutions, and others. Her research involves clinician-facing and patient-facing technology interventions to improve healthcare, particularly among minority, underserved, or otherwise non-mainstream patients. Her work at 9 academic institutions has put her in a position to describe lessons learned. She has acted as formal mentor in the NIH PRIDE program, University of Pittsburgh’s iBRIC and Downstate’s Arthur Ashe programs for research training for minority high school and undergraduate students.

Anticipated Audience

Informatics professionals seeking a deeper understanding of why a diverse and inclusive workforce and environment are crucially important goals for the informatics field. Informatics professionals looking to learn about best practices and resources for broadening participation in informatics.

Discussion Questions

1. How can we ensure that we are not perpetuating the myth of the pathway problem in informatics?
2. What concrete steps can each person take within their own organization to move towards a more inclusive
3. What are some ways organizations can ensure persons historically excluded from health informatics positions of influence based on ethnicity or race (PEERs) are mentored and positioned for career advancement opportunities?

Participant Statement
All panelists have agreed to take part in the panel.

Conclusion
Cultivating a diverse and inclusive workforce is critical for the biomedical informatics profession to advance health and healthcare transformation; this is a foundational step to ensuring informatics-enabled solutions are equity-centered and deliver equitable patient and population outcomes. To achieve these goals, intentional and deliberate action is needed to address systemic inequalities that have resulted in exclusions of individuals due to ethnic or racial background (PEERs) in STEM fields such as biomedical informatics.

References
Use of Healthcare Information Technology and Data Platforms to Inform Pandemic Response: Perspectives at the Intersection of Academic Healthcare Provider Organizations and Industry Partners

Atul J. Butte, MD, PhD1; Christopher A. Longhurst, MD, MS2; Amy P. Abernethy, MD, PhD3; Gretchen P. Jackson, MD, PhD4,5; Philip R.O. Payne, PhD6

1University of California San Francisco, Bakar Computational Health Sciences Institute, San Francisco, CA; 2University of California San Diego, Departments of Biomedical Informatics and Pediatrics, La Jolla, CA; 3Verily, San Francisco, CA; 4IBM Watson Health, Cambridge, MA; 5Vanderbilt University Medical Center, Departments of Surgery, Pediatrics, and Biomedical Informatics, Nashville, TN; 6Washington University in St. Louis, Institute for Informatics, St. Louis, MO

Abstract
This panel will summarize the collective experience of several academic healthcare provider organizations and industry partners as those organizations sought to understand and respond to the COVID-19 pandemic in the US. Such measures included collecting and analyzing data corresponding to healthcare organizations, public health departments, socioeconomic indicators, and additional signals collected directly from individuals and communities. In addition, our panelists will describe their experiences when seeking to use electronic health record (EHR) derived data, given the critical role of EHRs when seeking to improve clinical care, research and inform public health decision-making. The speakers will also describe their perspective on current challenges in the data ecosystem and the technology infrastructure relevant to the COVID-19 response. Such infrastructure includes registries and clinical data networks to support population-level analyses. Finally, we will propose a specific set of strategic next steps to increase interoperability, overall organization, and efficiencies of such technologies and approaches, improve the delivery of care at the individual and population levels, and prepare for future public health emergencies.

Introduction
Healthcare information technology (HIT) and the data assets derived from such platforms have and continue to be essential to our collective response to the COVID-19 pandemic [1, 2]. The use of such technology and data can inform clinical decision-making [3, 4], enable resource and personnel management [5], and enhance inter-and intra-organizational communications and coordination. Realizing these types of benefits, given the unique challenges presented by the pandemic, have required close collaboration between healthcare provider organizations and industry, as well as engagement by diverse experts, including information technology professionals, informaticians, and data scientists, to name a few of many such roles and viewpoints. These multi-disciplinary collaborations and teams have served to change our collective thinking as to how such entities and individuals can and should interact and how those interactions, in turn, can improve the quality, safety, and outcomes of care, even in emergencies. As such, they are informative not only to how we respond to COVID-19 in the present but also how we re-envision the HIT and data infrastructure and operations of a modern healthcare system that can deliver improved quality, safety, and outcomes of care, in the context of both conventional care delivery, as well as future public health emergencies.

This panel will review the experiences of a group of academically based healthcare provider organizations and industry partners, working in close coordination with regional partners and communities of practice to respond to COVID-19, leveraging existing HIT and data platforms and resources. We will then go on to propose a set of future research and practice directions, informed by those experiences, to ensure we learn from this collective set of “lessons learned” to improve our ability to translate biomedical informatics and data science innovations into widespread use, via the creation and sustainment of cross-sector and stakeholder collaborations.

Topic importance and timeliness
The current challenges and opportunities surrounding HIT and associated data platforms to respond to COVID-19 are compared to previous pandemics, such as SARS and Ebola; there are essential differences in said experiences. These differences are primarily a function of 1) broad adoption and use of electronic health records (EHRs) and associated clinical decision support and analytical capabilities; 2) rapid advances in biomedical data science, spanning a spectrum
from data integration and harmonization to visualization to the application of Artificial Intelligence (AI); and 3) increasing availability of socio-demographic, patient-generated, and sensing data that extend beyond the traditional clinical environment. These new resources and capabilities can and should enable more rapid translation of biomedical informatics and data science innovations from the lab to operational use, as seen in our response to COVID-19. However, a shared “roadmap” for how such “informatics translation” can and should occur does not exist, which may preclude the full realization of the benefits of moving innovative solutions into widespread use. This represents a significant gap in knowledge and practice that has surfaced as a function of the current pandemic and that we must continue to address as we move into our “new normal.” Suppose we are successful in addressing this gap. In that case, as is noted above, we can leverage HIT and data assets in new and innovative ways that can improve conventional care delivery at the individual and population levels and our preparedness for future public health emergencies that may present themselves. This panel will focus on the above issues, as it is intended to catalyze a robust, community dialogue concerning how we can and should evolve the intersection of biomedical informatics and data science with both healthcare provider organizations and industry partners to improve “informatics translational” and ensuing benefits to the patients and communities we collectively serve.

Speakers

Dr. Atul J. Butte (University of California San Francisco)

Atul Butte, MD, PhD is the Priscilla Chan and Mark Zuckerberg Distinguished Professor and inaugural Director of the Bakar Computational Health Sciences Institute (bchsi.ucsf.edu) at the University of California, San Francisco (UCSF). Dr. Butte is also the Chief Data Scientist for the entire University of California Health System, the tenth largest by revenue in the United States, with 20 health professional schools, 6 medical schools, 6 academic health centers, 10 hospitals, and over 1000 care delivery sites.

Dr. Christopher A. Longhurst (University of California San Diego)

Dr. Longhurst serves as Chief Information Officer and Associate Chief Medical Officer for Quality and Safety at UC San Diego Health. As CIO, Dr. Longhurst oversees all operations and strategic planning for information and communications technology. He is responsible for planning and developing all administrative and clinical information systems related to operating UC San Diego Health hospital and clinical facilities, including electronic health records and the MyUCSDChart system. As Associate CMO for Quality and Safety, Dr. Longhurst is focused on building upon current quality improvement efforts and aligning people, processes and technology to achieve UC San Diego Health’s quality and accountability goals.

Dr. Amy P. Abernethy (Verily)

Dr. Abernethy is the President of the Clinical Research Business at Verily, an Alphabet Company founded at the convergence of healthcare, data science and technology, where she has responsibility for the company’s Baseline program and other initiatives to support a broad range of clinical trials and real-world evidence (RWE) studies. Before joining Verily in July, Dr. Abernethy was Principal Deputy Commissioner and Acting Chief Information Officer of the US Food & Drug Administration.

Dr. Gretchen P. Jackson (IBM Watson Health and Vanderbilt University Medical Center)

Dr. Gretchen Purcell Jackson is the Vice President, Chief Health and Science Officer at IBM Watson Health and an Associate Professor of Surgery, Pediatrics, and Biomedical Informatics at the Vanderbilt University Medical Center. She is an internationally recognized informatician and accomplished clinical surgeon with over 25 years of contributions to informatics research, innovations in health information technologies, and surgical science. She is chair-elect for the Board of Directors for the American Medical Informatics Association (AMIA) and an elected fellow of the American College of Medical Informatics (ACMI).

Dr. Philip Payne (Washington University in St. Louis)

Dr. Payne is the Associate Dean for Health Information and Data Science and Chief Data Scientist at the Washington University School of Medicine. In addition, he is the Janet and Bernard Becker Professor and founding Director of the school’s Institute for Informatics (I2). Since March of 2020, Dr. Payne has led the joint COVID-19 analytics workgroup convened by Washington University and its partners at BJC Healthcare. This workgroup has served to not only support the operational needs of Washington University Physicians and BJC Healthcare in meeting the challenges posed by COVID-19 but has also served as the primary data, information, and knowledge “engine” for the St. Louis
Metropolitan Pandemic Task Force, a coalition of elected officials, public health organizations, and HPOs that collectively serve the 3.4M citizens of the St. Louis MSA.

Questions to be addressed:

1) How have cross-organizational and sector partnerships helped to accelerate “informatics translation” during the COVID-19 pandemic, and how can we replicate those outcomes as we enter into a post-pandemic “new normal”?

2) What gaps in current HIT and data platforms have the pandemic surfaced that require further investigation or the development of novel solutions? Similarly, where have we demonstrated that the maturity of HIT and data platforms is sufficient such that further research is not necessarily needed or likely to generate substantial impact?

3) How can healthcare provider organizations work with industry more effectively to accelerate novel HIT development, deployment, and scaling, in more “normative” circumstances than the current public health emergency?

4) How should we prepare both a research and development roadmap and workforce development strategies to ensure that the Biomedical Informatics and Data Science communities are actively engaged in and providing value to these types of efforts?

Conclusion

Our group of panelists represents the type of cross-organizational and sector partnerships that have both enabled successful responses to COVID-19, as that can serve as a template for accelerating future “informatics translation” when we move into a new (non-emergent) normal. By engaging the AMIA community in a dialogue surrounding these issues, we hope to promote the development of a shared “roadmap” that will ensure the success of the impact of Biomedical Informatics and Data Science research and practice in such contexts, both improving conventional care delivery, as well as ensuring we are better prepared for future, analogous public health emergencies.

References


LEAP 2020: Cutting Edge Health Information Technology Tools for Research

Kevin Chaney, MGS1, Kenneth D. Mandl, MD, MPH2, Kristen Miller, DrPh, CPPS3,
1Office of the National Coordinator for Health Information Technology, Washington, DC;
2Computational Health Informatics Program, Boston Children’s Hospital, Boston, MA;
3National Center for Human Factors in Healthcare, MedStar Health, Washington, DC

Abstract

The utilization of application programming interfaces (APIs) in healthcare has potential to enhance population health, patient care and research. Furthered by regulation issued by the Office of the National Coordinator for Health Information Technology (ONC) and the Centers for Medicare & Medicaid Services (CMS), patients and providers will have greater access to electronic health information. In anticipation of a new generation of health IT, ONC issued the Leading Edge Acceleration Projects (LEAP) in Health IT funding opportunity, to advance well-designed, interoperable, and scalable health IT for care and research. This panel will present findings, learnings, and accomplishments from the LEAP work to-date and discuss objectives of the newest 2020 funded initiatives, focused on reducing provider and health system burden of utilizing health IT, while incorporating data access and use via an API for innovative purposes in support of care and research. The panel will showcase the two most recent innovative initiatives, an effort to build a platform that leverages bulk data to support an ecosystem for research and learning, and a landscape analysis to explore, identify, and describe current open health IT-based tools for research.

Introduction

The Office of the National Coordinator for Health Information Technology (ONC) is at the forefront of the administration’s health information technology (IT) efforts and leads efforts to support the adoption of health information technology and promotion of nationwide health information exchange to improve healthcare. Over the last decade, ONC has made tremendous progress towards advancing not only the adoption and use of health IT across the healthcare ecosystem, but also increasing the rate of growth and innovation through its many programs to bridge policy with operational initiatives that can accomplish its mission.1 Programs such as the Strategic Health Information Technology Advanced Research Projects (SHARP) cooperative agreements2 aimed to close the gap between the promise of health IT and its realized benefits. Hailed as the major achievement from the SHARP program was the Substitutable Medical Applications, Reusable Technologies (SMART) Health IT,3 a standards-based technology platform that enables innovators to create medical applications (apps) that seamlessly and securely run across the healthcare system.4,6

More recently, there has been a rapid progression in the types of technologies and innovations being utilized in all health domains, including personal use for patient generated health data, clinical tools and apps being used at the point of care, and surveillance and population health tools that may include social determinants of health data. As the electronic exchange of health data has matured, the amount and types of health data available has expanded as well. Data standards such as Health Level Seven International’s (HL7®) Fast Healthcare Interoperability Resources (FHIR®) and application programming interfaces (APIs) are facilitating the sharing of health data in an interoperable way while using open standards that even non-healthcare industry technologists can leverage. Despite these advances, there is still much to learn about delivering and presenting emergent data seamlessly to patients and providers in both clinical and non-clinical settings in support of care and research. To better address these gaps and support alignment between the clinical and research ecosystems, ONC released National Health IT Priorities for Research: A Policy and Development Agenda (the Agenda),3 which identifies nine priority areas and corresponding actions. ONC’s is working to advance these priorities through a variety of actionable programs and policies aimed at: leveraging EHR data to support patient- and population-level research, analyses and services; improving patient engagement applications; enhancing consent management platforms; and improving tools to integrate clinical knowledge into routine clinical practice.

Leading Edge Acceleration Projects (LEAP) in Health IT

In 2018, ONC published the Leading Edge Acceleration Projects (LEAP)3 in Health IT funding opportunity. The goal of this three-year funding opportunity is to further a new generation of health IT development and inform the innovative implementation and refinement of standards, methods, and techniques for overcoming major barriers and challenges in the field. This panel presents in detail the focus of LEAP in Health IT funding opportunity and the
ongoing work of the two 2020 LEAP funded projects for Special Emphasis Notice Area of Interest #2: Cutting Edge Health IT Tools for Scaling Health Research. Key aspects of these projects focus on reducing provider and health system burden of utilizing health IT, while incorporating data access and use via an API for innovative purposes in support of care and research. LEAP in Health IT is an example of ONC’s ability to fund projects that seek to overcome challenges that inhibit the development, use, or advancement of well-designed, interoperable health IT affecting care and research.

Panel Objectives and Presenters
This panel will provide an overview of how projects funded under the ONC’s LEAP in Health IT funding opportunity is preparing the U.S. for an interoperable, modular, health IT ecosystem. Panelists will discuss results from projects awarded in 2020, in addition to current and emerging challenges facing the field, results of canonical use case prototypes, and priorities for next steps in alignment with national health IT priorities for research. Each panelist will leave time for discussion around the development, implementation, and use of their respective solution, as well as early legal and policy implications. This panel aims to stimulate a highly interactive discussion and strengthen the community’s knowledge of innovative uses of health IT tools, barriers and solutions for use, and areas ripe for future work.

Mr. Kevin Chaney (moderator and organizer), is a Senior Program Manager at ONC and co-leads ONC’s the LEAP in Health IT program. He will introduce and moderate the session, provide an overview of ONC’s relevant portfolio of work and describe the goals of the LEAP in Health IT program. Mr. Chaney will facilitate discussion with panelists on current use of LEAP in Health IT-funded novel technologies beyond leading-edge health organizations and explore opportunities for mainstream health systems without academic or research affiliations.

Dr. Kenneth D. Mandl (panelist), directs the Computational Health Informatics Program at Boston Children’s where he leads the transformative SMART Health IT initiative and is Principal Investigator of a 2020 ONC LEAP in Health IT project. He will provide an overview of how the SMART/HL7 Bulk Data (Flat FHIR) API has been used to provision data at scale through a cloud-based platform called Cumulus, which can support an apps ecosystem for research and learning. Dr. Mandl is building a health IT ecosystem underpinned by standardized, public APIs. The project relies on the emerging SMART/HL7® FHIR® bulk data export standard, which has already been implemented in EHRs through the Argonaut process, and is an HL7 Standard for Trial Use (STU).

Dr. Kristen Miller (panelist), is the Scientific Director of the National Center for Human Factors in Healthcare at MedStar Health, an Associate Professor of Emergency Medicine at Georgetown University School of Medicine, and Associate Faculty at the Innovation Center for Biomedical Informatics at Georgetown Medical Center. She is the Principal Investigator of a 2020 ONC LEAP in Health IT project. She will provide an overview of MedStar Health’s continued work to advance knowledge at the point of care. She will discuss MedStar’s FHIR Factories, which represent new technology approaches in big data automation to better align the clinical and research ecosystems so researchers can pursue more complex research questions and make faster, more reliable discoveries.

Panel Discussion Questions
- How are advancements from these projects available for other organizations to use and leverage?
- Are there disruptive technologies the field should be prepared for?
- What needs or gaps have these new open health IT-based tools addressed and what remains?
- What technical and policy needs or gaps limit the usage of open health IT-based tools for health care and research?
- Are there health IT infrastructure and/or standards barriers impeding the utilization of these tools?
- How easily can these tools be applied and scaled to other similar aspects in the field (e.g., other risk calculators, bulk data types, consent resources, or patient-engagement technologies)?
- How can ONC broaden the pool of use cases able to utilize these established technologies, standards, and tools (e.g., linking and aggregating research-relevant data sources; improving the ability to match individuals to different data sources; supporting robust de-identification to increase confidence and manage risk)?
- How does ONC ensure that new technologies and tools improve access and use of health information for care and research?

Panel Learning Objectives
1. Participants will understand ONC’s priorities, with a strong focus on the LEAP in Health IT Program, and the newest focal areas of interest.

2. Participants will learn about the technical development and utilization of each LEAP in Health IT project to date.

3. Participants will learn the challenges and barriers experienced by each LEAP in Health IT project and the impact for broader uptake by the field.

4. Participants will learn about technical and policy gaps that are limiting the use of open health IT-based tools.

Conclusion

This panel will discuss the rapid progression of innovative health IT solutions and approaches demonstrated and forthcoming as part of ONC’s LEAP in Health IT initiative. This panel aims to stimulate a rich discussion and gather participant input that will inform and strengthen the innovative work being explored. Discussion will also include the current and future needs and actions required to address emerging technical challenges and policy implications from multiple perspectives.

Statement of Participation

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

References


Addressing the Digital Divide to Promote Health Equity

Guilherme Del Fiol, MD, PhD1, Chelsey R. Schlechter, PhD, MPH1, Bryan Gibson, DPT, PhD1, Thomas J. Reese, PharmD, PhD2, David W. Wetter, PhD1

1University of Utah, Salt Lake City, UT; 2Vanderbilt University, Nashville, TN

Abstract

Despite wide adoption of electronic health records (EHRs), a “digital divide” persists between high- and low-resource healthcare settings. However, informatics interventions that specifically target underserved populations can reduce inequities. In this panel, a multidisciplinary group of experts will discuss aspects of a research-practice partnership across multiple federally funded projects that aim to increase the reach and uptake of evidence-based prevention and treatments in underserved populations. Presentations will cover 1) examples of pragmatic clinical trials focused on informatics interventions to increase the uptake of evidence-based prevention and treatments, along with a multisector partnership that supports these implementation studies; 2) rapid-cycle design and adoption of informatics interventions; 3) sociotechnical and EHR adaptation approaches; and 4) population health management approaches. Presenters will cover a broad range of projects, including a study case focused on increasing the uptake of COVID-19 vaccination and testing in underserved communities.

Introduction

Despite wide-spread adoption of electronic health record (EHR) systems among low-resource healthcare settings, a digital divide persists between high-resource systems and low-resource settings such as Community Health Centers (CHCs).1 Rural and underserved settings are ~40% less likely to have comprehensive EHRs; 17% less likely to allow patients to view, download, and transmit their health information; less likely to have adopted patient engagement functions; and have lower digital patient-provider communication.1-3

Panelists will describe a state-wide, multisector partnership that includes multiple pragmatic implementation studies investigating informatics interventions designed for transformative impact, sustainability, and dissemination. This research-practice partnership addresses health inequities and the growing digital divide between high-resource healthcare systems and socioeconomically disadvantaged settings. Informatics interventions include EHR-based clinical decision support (CDS) tools and population health management. Implementation barriers are addressed via sociotechnical assessment of barriers and facilitators; user-centered design; and continuous support throughout implementation and testing. The partnership leverages: 1) ubiquitous information technology (e.g., capabilities required for EHR certification, text-messaging for patient outreach) to address the lack of resources that plague low-resource settings and underserved populations; 2) multilevel interventions (e.g., clinic and patient-levels); 3) sustainable, evidence-based interventions that have demonstrated efficacy among underserved populations; and 4) multisector partners (i.e., CHCs, Primary Care Association, State Department of Health, Academic Institution).

The panel includes a multidisciplinary team of speakers with expertise in health equity, health behaviors, implementation science, sociotechnical methods, and clinical informatics. Each of the speakers will describe a critical aspect of the research-practice partnership with CHCs to address the digital divide: 1) pragmatic clinical trials focused on informatics interventions to increase the uptake of evidence-based prevention and treatments, along with a multisector partnership that supports these implementation studies (Wetter); 2) rapid-cycle design and adoption of informatics interventions (Schlechter); 3) sociotechnical and EHR adaptation approaches (Gibson); and 4) population health management (PHM) approaches (Reese).

Guilherme Del Fiol is an Associate Professor and Vice-Chair for Research at the University of Utah Department of Biomedical Informatics, and a co-chair of the Health Level Seven (HL7) CDS Work Group. He has over 20 years of experience investigating clinical informatics interventions to improve the quality, efficiency, and value of healthcare, with a primary focus on promoting health equity. Dr. Del Fiol will introduce the panel and moderate the session.

David W. Wetter is the Jon M. and Karen Huntsman Presidential Professor at the University of Utah and Huntsman Cancer Institute, Director of the Center for Health Outcomes and Population Equity (HOPE), and Associate Director for Practice Engagement and Translation at the Clinical and Translational Sciences Institute. His work is targeted at eliminating inequities related to chronic and infectious disease through translational research, with a major focus on low socioeconomic status, rural/frontier, and diverse groups.
Dr. Wetter will present examples of pragmatic clinical trials led by the Center for HOPE with a focus on multi-level informatics interventions to increase the uptake of evidence-based prevention and treatments, as well as multisector partnerships that support these implementation studies. Implementation of evidence-based interventions can be challenging and often requires coordination across stakeholders from multiple levels within an organization (e.g., healthcare system leadership, clinicians, patients) and across organizations (e.g., healthcare systems, public health agencies, academic institutions). Community-engaged dissemination and implementation research emphasizes engaging stakeholders (e.g., community members, practitioners, community organizations) with diverse perspectives, and experience to provide tacit knowledge regarding the local context and priorities. The aims of this presentation are three-fold. 1) Describe a cross-sectoral, state-wide, research practice-partnership infrastructure that was developed to improve health inequities. The partnership consists of the state Primary Care Association, CHC systems, the State Department of Health, and an Academic Research Institution. 2) Describe the community-engagement process for the projects undertaken by the partnership. Each project incorporates ubiquitous health information technology at the level of the CHC system and/or the patient and are designed to implement evidence-based interventions to address infectious and chronic disease prevention and control (i.e., tobacco cessation, HPV vaccination, colorectal cancer screening, COVID-19 testing and vaccination). 3) Describe the contribution of the partnership and engagement activities to the development, adaptation, and implementation of health information technology implementation strategies. This partnership and community-engagement process depicts a tangible and replicable example of how expertise, resources, and data can be shared across multisector partners to ultimately capitalize on assets of each partner to accomplish a shared goal of improving population health and eliminating health inequities.

Chelsey R. Schlechter is an Assistant Professor in the Center for HOPE and the Department of Population Health Sciences at the Huntsman Cancer Institute and University of Utah. Dr. Schlechter is an Implementation Scientist whose research focuses on improving population health through the implementation and dissemination of evidence-based interventions for chronic disease prevention and control in community and healthcare settings.

Dr. Schlechter will present the rapid-cycle design and exemplar adaptations for an NIH-funded Radx-UP initiative designed to increase COVID-19 testing and vaccination among underserved populations. Developing, implementing, and adapting digital health interventions to address the evolving COVID-19 pandemic has required a rapid and iterative process. While still allowing rigorous evaluation, rapid-cycle designs enable iterative adaptation of interventions by incorporating data from multiple sources, including policy (e.g., vaccination approvals, eligibility), setting capacity (e.g., clinic testing capacity, vaccine availability), and stakeholder priorities (e.g., infection ‘hot-spots’ for testing/vaccination). The objectives of the presentation are to describe the application of a rapid-cycle design and adaptation process to digital health interventions; provide exemplars of adaptations for an ongoing trial to address COVID-19 among CHC patients (SCALE-UP Utah); and describe how implementation science frameworks can be used to design and evaluate adaptations to digital health interventions.

Bryan Gibson is an Assistant professor of Biomedical Informatics at the University of Utah. Dr. Gibson's expertise is in sociotechnical assessment methods and EHR adaptation to promote uptake of evidence-based interventions in underserved populations.

Dr. Gibson will present 1) sociotechnical methods used to guide the design and implementation of informatics interventions in low-resource settings, including group discussions with CHC leaders, direct observations of clinical workflow and EHR features, surveys of clinical staff, and assessment of EHR capabilities and implementation status at different CHCs; 2) examples of EHR-based tools connecting patients to evidence-based prevention and treatment services, such as CDS reminders coupled with electronic referrals to evidence-based prevention and treatment services; and 3) a study case focused on QuitSMART Utah, a pragmatic, comparative-effectiveness trial investigating the effect of an EHR-based intervention called Ask-Advise-Connect (AAC) on patient enrollment in tobacco cessation treatment. AAC consists of systematically assessing every patient for tobacco use during the visit intake (Ask), offering a connection to a tobacco cessation quitline (Advise), and submitting an electronic referral to the quitline (Connect). AAC has been implemented at 11 CHCs and 39 clinics across Utah using three different EHR products. Dr. Gibson will also explain how sociotechnical methods have been adapted to overcome constrains imposed by the COVID-19 pandemic to support the uptake of evidence-based interventions such as HPV vaccination, colorectal cancer screening, and COVID testing and vaccination.

Thomas J Reese is an Assistant Professor in the Department of Biomedical Informatics at Vanderbilt University, a Board-Certified Ambulatory Care Pharmacist, and a recent fellow in the National Cancer Institute’s Multilevel Training Institute. Dr. Reese serves on the editorial board for JAMIA and as the vice-chair of the AMIA
Dr. Reese’s research is focused on using health information technology to implement evidence-based practice in contextually and theoretically informed ways.

Dr. Reese will describe PHM interventions to increase the uptake of enrollment in tobacco cessation treatment, colorectal cancer screening, HPV immunization, and COVID-19 testing and vaccination among patients from 11 CHC systems comprised of 39 clinics throughout Utah. PHM approaches coupled with ubiquitous technology strategies, such as text messaging, have the potential to reduce health inequities by connecting patients to evidence-based interventions. However, low-resource settings often lack technical resources and expertise required to implement PHM strategies at scale. Multisector partnerships can help lower barriers to the implementation of PHM interventions at low-resource settings. Dr. Reese will describe components of PHM-based interventions, including 1) automatic identification of eligible patient cohorts based on EHR data; 2) automatic patient outreach through HIPAA-compliant text messaging that offers eligible patients a direct connection to vaccination and/or testing; and 3) patient navigation to help address barriers, hesitancy, and logistics. He will also describe rapid PHM interventions to increase the uptake of COVID-19 testing and vaccination.

Relevance to the AMIA 2021 Annual Symposium, Significance of the Topic, and Anticipated Audience

The proposed panel aligns closely with two themes of the symposium’s Late Breaking Submission:

1) Diversity, Equity, and Inclusion and the role of informatics. Reviews of public health programs have concluded that non-targeted evidence-based interventions for prevention may negatively impact health equity, whereas programs that specifically target underserved populations can reduce inequities.\(^4\) In this panel, presenters will share their experience with informatics interventions that are deliberately designed and implemented to reduce health inequities.

2) Informatics response to COVID-19: The COVID-19 pandemic has both exposed and exacerbated health inequities. Informatics has played a critical role in supporting health care organizations address the pandemic, such as in the substantial and rapid expansion of telehealth services. In this panel, presenters will share their experience with the rapid design and implementation of informatics interventions to help reduce health disparities related to COVID-19 testing and vaccination.

The anticipated audience includes individuals interested in clinical informatics and implementation research focused on promoting health equity in rural, low socioeconomic status, and racial/ethnic minority populations.

Discussion Questions for Audience Participation

- How can rapid-cycle designs improve implementation and effectiveness of digital health interventions?
- How can implementation science frameworks be used to benefit implementation of digital health interventions?
- What are key challenges to implementing informatics interventions across multiple EHR systems and healthcare organizations with low technical resources?
- How can population health management approaches be implemented at scale, minimizing dependencies on local technical resources and without increasing clinician workload?

Participation statement

All panelists have agreed to take part in the panel.

References

Reliable insights regarding medication outcomes from real world data—key principles of an evidence generation framework

Panelists: Sebastian Schneeweiss, MD, ScD1 Keith Marsolo, PhD2 Shirley Wang, PhD1 Robert Ball, MD, MPH, ScM3 Moderator: Rishi J Desai, PhD1

1 Division of Pharmacoepidemiology and Pharmacoeconomics, Brigham and Women’s Hospital and Harvard Medical School, Boston, MA
2 Department of Population Health Sciences, Duke University, Durham, NC
3 Food and Drugs Administration, Silver Springs, MD

Abstract (144/150)

Availability of richer clinical information on a wide-scale from electronic health records (EHRs), often stored in unstructured form, and advances in informatics methodology have opened up new opportunities to address questions related to medication safety and effectiveness that are unanswerable with structured information from other sources such as administrative insurance claims. However, generating decision-grade evidence from real-world data, including EHRs and insurance claims, requires a robust framework rooted in state-of-the-art principles of causal study design and analysis. This session will identify challenges and opportunities for addressing medication safety and effectiveness questions using real-world data sources. At the conclusion of this session, participants will be able to recognize the importance of robust design principles, careful data curation, and transparent reporting practices in deriving reliable insights regarding medication safety and effectiveness from real world data to inform regulatory actions or clinical practice recommendations.

Description of the panel

The Food and Drug Administration (FDA)’s Sentinel System uses distributed analytic tools and curated real-world longitudinal health insurance claims data for more than 100 million people from participating healthcare systems to generate insights regarding the safety of medical products. The FDA Sentinel Innovation Center was launched in 2019 to create a Medical Data Enterprise to enhance the Sentinel infrastructure by incorporating EHR data from at least 10 million lives. In this panel, Dr. Rishi J Desai, who is the Operations Chief of the FDA Sentinel Innovation Center, will first introduce the challenges in creating a national infrastructure comprising of various data streams including health insurance claims and linked EHRs that can enable timely assessments of pressing safety questions in a moderated discussion section (15 minutes). Following the introductory remarks, panelists will provide a detailed overview of ongoing activities to address these challenges in the following 4 sessions with active audience participation at the end of each session.

1. Causal inference principles guiding non-randomized studies of medication outcomes, by Dr. Sebastian Schneeweiss (20 minutes): This session will cover aspects related to design and analytic principles to allow emulation of
target trials from non-randomized data. Parameters including choice of the study design, choice of appropriate time horizon for the study to avoid time-related biases, clearly identified eligibility and observability based on the data available, and sensitivity analyses to quantify the impact of various design choices on study results will be covered. Dr. Schneeweiss will describe learnings from the RCT-DUPLICATE initiative,¹ which aims to evaluate the circumstances in which non-randomized studies conducted using real-world data can generate reliable causal inference.

2. Data curation in EHRs to maximize their value for clinical research, by Dr. Keith Marsolo (20 minutes): EHR data sources are heterogeneous in their content, structure, completeness, and quality. While a common data model can help to impose a standard organization of the data across multiple sites, mapping of source values into the model can lead to omissions, data errors, and other data quality issues. Moreover, even though a data model standardizes the data elements, a lack of semantic interoperability across sites may remain. Dr. Marsolo will describe data curation initiatives in the National Patient-Centered Clinical Research Network (PCORnet®) data infrastructure to assure data quality and reduce variability of information across participating sites.²

3. Transparency initiatives, by Dr. Shirley Wang (20 minutes): Transparent communication of study specifications in a pre-registered protocol, akin to a trial protocol, is a key feature that could help with increasing confidence in results from the non-randomized investigations. Dr. Wang will describe the STaRT-RWE reporting tool,³ which is developed as an instrument for designing and conducting reproducible studies using real world data; set clear expectations for transparent communication of methods; reduce misinterpretation of prose that lacks specificity; allow reviewers to quickly orient and find key information; and facilitate reproducibility, validity assessment, and evidence synthesis. Additionally, Dr. Wang will also cover principles guiding the current practices for quality control and transparent sharing of analytic codes in the Sentinel system.

4. A regulator’s perspective on evidence generated from real-world data, by Dr. Robert Ball (15 minutes): The FDA and other regulatory agencies around the world routinely face the challenging task of considering regulatory action based on results from non-randomized investigations of medication outcomes conducted using routinely collected healthcare data. Generating valid inference from non-randomized studies often using secondary healthcare data is complex owing to multiple potential sources of bias. Dr. Ball will outline the key considerations for evaluating results from non-randomized studies from a regulator’s perspective.

As regulators and policymakers move towards optimizing decisions impacting health of populations using evidence generated from real-world data, the issues discussed in this session are timely and ripe for wider dissemination and discussion in the clinical research and informatics communities. Ideas discussed in this panel will lead to a
stimulating exchange with the audience regarding the current and future role of informatic research community in addressing the unmet research needs.

A list of discussion questions to enhance audience participation

1. What are some key features of a robust evidence generation framework utilizing real-world data?
2. When is evidence generated from real-world data good enough to recommend regulatory or policy decisions?
3. What are the key informatics innovations that could facilitate broader and more efficient use of EHRs in medication safety and effectiveness research?
4. What are some key pitfalls to avoid when interpreting medication safety and effectiveness evidence generated from real-world data?

By the end of the session, the audience can expect to gain a deeper understanding of the challenging task of creating and maintaining a national infrastructure to enable efficient and reliable evidence generation regarding safety of medical products using real-world data. Further, this session will provide participants with a greater appreciation of the multidisciplinary nature of the evidence generation framework which requires application of principles from various disciplines including medicine, epidemiology, biostatistics, and informatics.

Statement:

Dr. Rishi Desai confirms that all panelists have reviewed the proposal and agreed to participate in the proposed panel.

References

Mental Functioning Ontologies

Guy Divita, MS¹, Maryanne Sacco MA¹, Janna Hastings, PhD²,³, Piper Ranallo PhD⁴

¹Rehabilitation Medicine Department, National Institutes of Health Clinical Center, Bethesda, Maryland, USA; ²Department of Clinical, Educational and Health Psychology, University College London, UK; ³Department of Computer Science, Otto-von-Guericke University Magdeburg, Germany; ⁴Six Aims for Behavioral Health, Minneapolis MN

Abstract

Mental functioning is an important component of health and health care. However, documentation of mental functioning is included mostly within unstructured parts of the medical record, is sparse, and is hard to extract. In addition, the field lacks a clear definition of mental functioning. We aim to raise the visibility of this domain within the informatics community, to establish a common definition of mental functioning, and to push for standardized terminologies for mental functioning so that health care providers can document mental functioning more consistently and accurately. This exploratory panel will facilitate awareness of those engaged in creating and using ontologies related to mental functioning, illuminate the direction the community is heading, and elicit use cases for these ontologies.

Learning Objectives

1. Define mental functioning and differentiate key concepts and their relationships for ontology building.
2. Explain how mental functioning ontologies are addressing gaps in current terminologies and classifications representing mental functions and mental functioning such as in SNOMED and the WHO ICF.
3. Examine how mental functioning ontologies can be applied to real world case examples from clinical practice, behavioral medicine, and disability determination adjudication.

Introduction

Mental functioning is an inherently complex domain, spanning multiple disciplines, data types, descriptive levels, and approaches. This complexity has brought considerable challenges to efficient knowledge discovery, integration across disciplines, and translation into the clinic [1]. Different vocabularies and semantic frameworks are used across these different descriptive levels, and it is difficult to gain an integrated and actionable overview of what is known. Interoperable and interconnected ontologies offer a technological platform for synthesizing across perspectives and integration of evidence and data [2]. Ontologies also have the potential to improve the reproducibility and findability of data within disciplines, accelerating the process of scientific discovery and translation into the clinic [3]. This panel brings an international perspective on developments in building ontologies to represent mental functioning. We will discuss clarifying the scope and definition of mental functioning, identifying concepts and their relationships, and establishing standardized terminologies that reflect current science and that are grounded in use cases. The panelists come from divergent and complimentary backgrounds with experiences in mental health informatics, mental and behavioral terminologies, and clinical practice. In our presentations we will illustrate how our different perspectives and skills can work together to support a common goal of transforming existing knowledge resources towards constituting a jointly interoperable framework of mental functioning ontologies.

The Panelists

Guy Divita/Maryanne Sacco

We are developing a mental functioning ontology based on two theoretical foundations. The ontology of mental functioning builds upon the frameworks of functioning from the International Classification of Functioning, Disability and Health (ICF)[4] and Open System Theory [5]. An ecological perspective is incorporated that provides concepts and their relationships and a terminology that aligns with how an individual’s functioning in everyday life activities while interacting within their environment is lived and described. Our long-term goal is to use the ontology to find mental-functioning mentions within text using machine learning-based extraction and classification tools that we
subsequently build. A terminology is derived from the ontology which is used for concept extraction within clinical documents to support our use case.

![Ecological Ontology of Mental Functioning](image)

**Figure 1.** Ecological Ontology of Mental Functioning

**Janna Hastings, PhD**

In her contribution to this panel session, Dr. Hastings will discuss the OBO Foundry [6] strategy for ontology interoperability and present the Mental Functioning and Emotion ontologies and their context and interrelationships to other OBO Foundry ontologies. Dr. Hastings will discuss how ontology re-use enables downstream integration and translation and mention some evidence synthesis efforts in the domain of behavioral medicine that are re-using these semantic frameworks. For example, the Human Behaviour-Change Project[7] has developed an artificial intelligence system for extracting and automatically synthesizing evidence on the effectiveness of behavior-change interventions from the literature of randomized controlled trials, and this project re-uses parts of the Mental Functioning Ontology to systematically annotate mental-functioning-related determinants of behavior change.

![The Mental Functioning Ontology upper level aligned to BFO](image)

**Figure 2.** The Mental Functioning Ontology upper level aligned to BFO.[8]

**Piper Ranallo, PhD**

In her contribution to the panel session, Dr. Ranallo will address clinical applications for ontologies of mental function. She will discuss the work of the SNOMED International Mental and Behavioral Health Clinical Reference Group[9]
she Chairs, as they have worked to define mental functions, processes and states from both an ontological perspective and concrete clinical perspective. She will discuss linkages between the representation of mental functions in SNOMED and the NIMH’s Research Domain Criteria (RDoC) framework, and provide an overview of how clinical ontologies currently represent the relationship between mental functions and the psychometric instruments and tasks used to measure and quantify them. She will provide overview of the landscape of (and gaps in) existing terminologies and ontologies relative to psychometric assessment instruments and tasks used to measure mental function. She will emphasize the need to improve our understanding of the distinction between mental functions, processes and states and the value proposition of improving the representation in existing clinical terminologies such as SNOMED.

Why the topic is timely, urgent, needed, and Anticipated Audience
There are increasing efforts in the health informatics community to establish a more unified terminology of mental functioning to support information extraction to meet diverse needs of stakeholders interested in mental functioning. Greater clarity is needed to define mental functioning concepts and their relationships. The target audience includes researchers and stakeholders in mental functioning and mental health, as well as informaticians that are interested in ontologies and terminologies in general.

Discussion Questions
1. What are your plans for collaboration across these different ontology development efforts?
2. What is the role of person factors within a mental functioning ontology?
3. What aspects of mental functioning make it harder to represent than other clinically relevant information?
4. How does the existence of multiple, inconsistent theories of mental function impact development of a single, coherent ontology of mental functioning?
5. How is clinical data captured and encoded in an EHR or clinical research system?
6. How are clinical terminologies such as SNOMED relate to research-driven ontologies such as the Foundry collection?
7. What implications does a mental functioning ontology have for clinicians?
8. Are there additional perspectives to consider that would dissect mental functioning into different components? For instance, would a neurologist, social worker, priest, pathologist, coroner, police officer, have new and or different facets than those represented here?
9. Can you offer some guidance/solutions to the problem of multiple, inconsistent theories of mental function?

Participation Statement
Guy Divita, Maryanne Sacco, Janna Hastings and Piper Ranallo assert that they agree to take part in this panel.

References
9. SNOMED International Mental and Behavioral Health Clinical Reference Group https://confluence.ihtsdotools.org/display/MBHCR
Addressing Bias in the Application of Machine Learning on Real-World Data

Hossein Estiri, PhD1,2, Yuan Luo, PhD3, Hongfang Liu, PhD4, Suzanne Tamang, PhD5,6,7, Harold Lehmann, MD, PhD8

1Lab of Computer Science, Massachusetts General Hospital, Boston, MA; 2Harvard Medical School, Boston, MA; 3Northwestern University Feinberg School of Medicine, Chicago, IL; 4Department of Artificial Intelligence and Informatics, Mayo Clinic, Rochester, MN; 5Stanford Center for Population Health Sciences, Stanford, CA; 6Department of Biomedical Data Science, Stanford University, Stanford, CA; 7Program Evaluation Resource Center, Office of Mental Health and Suicide Prevention, Department of Veterans Affairs; 8Johns Hopkins, Baltimore MD

Abstract

Applications of artificial intelligence (AI) and machine learning (ML) are transforming medicine towards an information processing discipline. Computer systems can now offer fast, cheap, and [in some cases more] precise diagnostic and prognostic decision support tools. Despite all the upside, AI/ML algorithms may also aggravate the various forms of biases that exist in today’s healthcare. This panel is designed to discuss the varied forms that algorithmic bias in AI/ML systems can take, including gender bias, racial prejudice, and age discrimination, across different areas of research in clinical informatics. The panel presentations will emphasize on lessons learned and informatics tools to measure and address bias in the application of AI/ML to real-world clinical data.

Introduction and Background

With the advent of large scale clinical data and the advancement of computational power, Machine Learning (ML) and Artificial Intelligence (AI) are increasingly applied to clinical data for knowledge discovery. “Bias” (“Systematic deviation of results or inferences from truth”) is a central concern in the analysis of observational data in general, and, therefore, clinical data, in particular. Such biases include, missingness, reliability, and manifestations of recording processes, access to healthcare, and patient interactions with the healthcare system[1]. AI/ML algorithms applied to these data can inadvertently be subject to such biases.[2,3] The fairness (or lack thereof) of resulting models is a concern that derives from methodological biases and that makes the models potentially cause social harm. Bias in analysis has been a concern of epidemiologists at least since Berkson’s bias was described[4] and since David Sackett made an initial list of potential biases in 1979.[5] In epidemiology, many methods for dealing with biases either in design or in analysis have been proposed.[6] At the same time that causal reasoning was returning to epidemiology, it was pursued by computer scientists as counterfactual reasoning, with many theorems and algorithms proved and designed to put such reasoning into practice.[7] The typical context has been regression-based analysis. With the popularity of AI and ML, those disciplines have now to implement in these novel contexts these age-old concerns.[8]

Panel Description

The panel will include a review of various biases that may be present in clinical data, and provide examples of how such biases are present in specific projects, how they were addressed, and successful the debiasing turned out to be. The panel will also discuss the historical and future implications of addressing (and not addressing) bias in AI/ML algorithms developed on clinical data.
Dr. Harold Lehmann will lead the panel discussion and will begin by setting the scene for the role of bias, and in doing so will help the audience better understand the importance of bias in the analysis of real-world data. This scene setting will include a taxonomy of biases that is applicable to all observational studies. Dr. Harold Lehmann is a Professor of Biomedical Informatics and of Medicine at the Johns Hopkins University School of Medicine.

**Presenters**

**Measuring Bias in Prediction of COVID-19 Outcomes (Estiri)**

Dr. Hossein Estiri is an Assistant Professor of Medicine at MGH lab of Computer Science and Harvard Medical School. In his presentation, Dr. Estiri will describe results from the measurement of various biases in modeling COVID-19 adverse outcomes using the MLHO framework,[9] including hospitalizat and ICU admissions, mechanical ventilation, and mortality. Clinical data often include more information for patients who have more clinical records -- i.e., more data. This is associated with demographic characteristics as a proxy for age, health state, and access to healthcare. AI/ML algorithms therefore can have inherent biases by providing more reliable predictions for patients with more clinical data. In his efforts designing the MLHO framework, we designed and implemented technical infrastructures to investigate bias. He will describe the results of investigating various biases (by demographics and healthcare utilization factors) in prospective evaluation of predictive models that utilize longitudinal (pre-Covid) medical records to predict Covid-related hospitalizat and ICU admissions, mechanical ventilation, and mortality. Finally, he will also discuss the informatics tools built in the MLHO framework to measure bias along with discrimination and reliability of binary classification algorithms.

**Mitigating Bias through People-Centered Health Data Science (Liu)**

Dr. Hongfang Liu is the Dr. Richard E. Emslander Professor of Biomedical Informatics at Mayo Clinic and a fellow of American College of Medical Informatics (FACMI). Her research interests include big data empowered analytics and informatics to facilitate real-world translation and implementation. She has committed to open science and has led the Open Health Natural Language Processing (OHNLP) effort which aims to develop and disseminate open source NLP tools to facilitate the use of real-world data for clinical and translational science research and health care delivery improvement. She has collaborated broadly with team science research experience in bioinformatics, software engineering, big data, cancer research, epidemiology, and clinical research. Dr. Liu will discuss the importance of people-centered AI/ML when conducting research in applying AI/ML techniques for real-world data. Drawing upon her experiences in translating NLP solutions to real-world use cases, she will discuss the importance of team science, open science, and citizen science to advance AI/ML for common good.

**Observing ML model biases in both EHR datasets and epidemiological datasets (Luo)**

Yuan Luo, PhD, is an Associate Professor at the Department of Preventive Medicine, Division of Health Biomedical Informatics at Feinberg School of Medicine in Northwestern University. He is Chief AI Officer at Northwestern University Clinical and Translational Sciences Institute and Institute for Augmented Intelligence in Medicine. His research interests include machine learning, natural language processing, time series analysis, computational phenotyping and integrative genomics, with a focus on biomedical applications. He won the American Medical Informatics Association (AMIA) New Investigator Award in 2020. He is currently an editor with JAMIA Open, JBI, Plos One, JHIR. He served on the AMIA Membership and Outreach Committee. Dr. Luo will discuss that ML models trained on datasets that lack demographic diversity could potentially yield sub-optimal performance when being applied to the underrepresented populations (ethnic minorities, lower social-economic status), thus perpetuating health disparity. Drawing from experiences of investigating this issue in both EHR datasets and epidemiological datasets, he will talk about disparities in ML model performances in racial, gender, and insurance subgroups. He will also discuss approaches to design models that proactively adjust for potential biases and to include subgroup reporting in their studies.
Racial and Ethnic Bias in Real World Risk Prediction  (Tamang)

Suzanne Roberta Tamang, PhD, is the Assistant Faculty Director of Data Science at the Stanford Center for Population Health Sciences and an Instructor in the Department of Biomedical Data Science, Stanford University. In addition to her academic role at Stanford, she is a Health Science Specialist/Statistician with the Office of Mental Health and Suicide Prevention, Department of Veterans Affairs, where she leads an inter-agency group charged with operationalizing natural language processing and information extraction tools within the Veterans Health Administration, for suicide and overdose prevention. Dr. Tamang will discuss recent work with the VA’s “REACH VET” suicide prevention program and the FDA’s Office of Minority Health and Health Equity on the assessment of racial and ethnic bias within real world risk prediction tools, using the VA’s stratification tool for opioid risk mitigation (STORM) as a use case. This includes a framework developed with Stanford Data Science for Social Good students to support joint work between data scientists and subject matter experts to quantify racial and ethnic bias presented by the deployment of advanced predictive analytics within mental health operations, for the purpose of developing strategies to ensure health equity.

Discussion Questions

1. What are the key biases that arise in the analysis of observational/real-world data?
2. How can different biases present themselves in AI/ML applied to clinical data?
3. How can bias be measured in AI/ML models developed on clinical data?
4. What are effective ways to model and correct for biases in clinical data?
5. What informatics tools are needed to help analysts address any biases in their data?
6. What tools are needed to help analysts convey detected biases in their models to non-technical subject matter experts, for the purpose of developing strategies to address such biases?

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

References

Transforming Healthcare through Patient-Generated Health Data Integration

Chun-Ju Hsiao, PhD1, Deborah Cohen, PhD2, Ida Sim, PhD, MD3, Leslie Lenert, MD4, Danielle Lavallee, PharmD PhD5
1Agency for Healthcare Research and Quality, Rockville, MD; 2Oregon Health & Science University, Portland, OR; 3University of California - San Francisco, San Francisco, CA; 4Medical University of South Carolina, Charleston, SC; 5British Columbia Academic Health Science Network, Vancouver, Canada

Abstract

With the advancement of digital technologies, patients can easily collect their own health data outside of the clinical setting to track and manage their health. While these technologies are promising, many ambulatory care practices lack the technical infrastructure, functional workflows, workforce capacity, and training to support the intake and use of patient-generated health data (PGHD). This panel will feature opportunities and recommendations regarding PGHD integration in ambulatory care settings. Specifically, the panel will feature findings from a recent environmental scan and guidance from a practical guide as well as examples from two health systems that use different technical approaches to collect and integrate PGHD with electronic health records. The learning objectives include: (1) understanding the current state and recommendations of PGHD integration, and (2) learning different technical architectures and workflows that support the collection and integration of standardized PGHD either self-reported by patients or directly from devices.

Panel Description

Patient-generated health data (PGHD) are “health-related data created, recorded, or gathered by or for patients (or family members or other caregivers) to help address a health concern.”[1] PGHD such as patient-reported outcomes (PRO), health-relevant behaviors (e.g., sedentary time, physical activity, sleep quality), anthropometrics (e.g., height, weight), and physiological measures (e.g., heart rate) can be collected through a variety of tools including patient portals, mobile applications (apps), and devices.

The potential for PGHD to impact health care delivery is significant. Ambulatory care clinicians base their decisions on information received from the patient, traditionally from data collected in the clinical setting. PGHD offers insights into the day-to-day health of an individual, providing patients and clinicians the ability to employ better strategies to prevent and manage acute and chronic conditions, through monitoring for remission and relapse. In addition, clinicians and scientists can use these data to generate and apply analytical techniques to improve risk prediction and diagnoses.[2]

The ongoing COVID-19 pandemic bolstered the need to collect and integrate PGHD with electronic health records (EHRs). Many ambulatory care face-to-face visits were replaced with virtual visits. While virtual care has the capability of ensuring patient access, virtual visits make it difficult to collect data essential for diagnosis and chronic disease monitoring. Clinicians now need to rely on patients to collect vitals and other health data that they previously collected in clinic. However, not all ambulatory care practices have the resources and capacity to support the intake of such data. Barriers include technical challenges related to accuracy of measurements, data provenance, interoperability, and privacy and security concerns in the data lifecycle (i.e., collection, transmission, storage, and analysis). Nevertheless, standardized interoperable data interfaces are quickly becoming important tools in the integration of PGHD into EHRs. These include standards such as Substitutable Medical Applications and Reusable Technologies (SMART) and Fast Healthcare Interoperability Resources (FHIR). Clinicians also face other challenges such as recommending valid and equitable devices from an increasing number of options, and prohibitive costs or inadequate information technology literacy for many patients.

This panel will share the current state of PGHD integration in ambulatory care settings from a recent environmental scan as well as recommendations from a PGHD integration guide. The guide provides considerations in various aspects including legal and compliance, technical architecture, operations, patient population, privacy, and security. The panel will also feature PGHD integration examples from two health systems that use different technical approaches (e.g., application programming interfaces, SMART on FHIR apps, CDS Hooks, and SMART Marker framework). Each of
the health systems developed different architectures to support the collection and integration of standardized PGHD either self-reported by patients or directly from devices. The presentations will be followed by a discussant-led conversation to stimulate thoughts about opportunities and best practices in collecting and integrating PGHD with EHRs, with a particular focus on how to work with different practices to achieve successful implementation and use of PGHD data for clinical care. The panelists represent different institutions and have been conducting research related to PGHD integration.

Chun-Ju Hsiao, PhD (organizer and moderator) is a health services researcher within the Digital Healthcare Research Division at the Agency for Healthcare Research and Quality (AHRQ). She leads AHRQ’s PRO initiative where she manages a portfolio of grants that aim to collect and use PRO data through innovative health information technology strategies. She also led a multi-agency project where AHRQ developed a new app and modified an existing app using the PRO FHIR Implementation Guide.[3] The two apps were pilot tested in ambulatory care practices and yielded valuable lessons learned in collecting and integrating standardized PRO data. Dr. Hsiao also oversees a project to develop a practical guide regarding PGHD implementation for ambulatory care practices. She will introduce the speakers and the topic including the rationale and motivation for bringing this panel together.

Deborah Cohen, PhD (panelist) is Professor of Family Medicine at the Oregon Health & Science University. Her areas of expertise include studying primary care innovation, with a focus on how health information technology is used and can be optimized in this setting. Dr. Cohen will present her team’s work on PGHD. She will highlight key findings from a systematic review of the literature on implementation of PGHD in ambulatory care settings, and describe the guide that she and her team are creating to help support ambulatory care leaders, at multiple levels, with decisions related to implementation of PGHD.

Ida Sim, MD, PhD, (panelist) is Professor of Medicine at the University of California, San Francisco (UCSF) and is Director of Digital Health for the Division of General Internal Medicine. She has worked in mobile health since the field’s earliest days, as Co-Founder of Open mHealth, a non-profit organization for open data exchange standards in mobile health. She has led multiple projects integrating PGHD into primary care at UCSF. Dr. Sim will present the panel’s first PGHD integration example. She will present on a generic PGHD architecture that is being used to integrate standardized PRO and sensor data. In one project that aims to improve outcomes among patients with multiple chronic conditions, self-reported PRO and PGHD data (e.g., depression, physical function, pain, blood pressure (BP), heart rate (HR), and oxygen saturation) are captured in SMART Marker format and made available as FHIR resources to a clinician-facing SMART-on-FHIR visualization app. Another project extends the same architecture to demonstrate device-agnostic (i.e., regardless of device model, make, location, and connectivity) integration of Bluetooth BP data. Using a cloud infrastructure based on FHIR, Open mHealth, and other standards, the system authenticates, ingests, harmonizes, and serves up BP and HR data from the BP cuffs as FHIR observation resources.

Leslie Lenert, MD (panelist) is Assistant Provost for Data Science and Informatics and the Chief Research Information Officer of the Medical University of South Carolina (MUSC). Dr. Lenert works as a “health systems engineer” to create advanced social technical processes (including technology-based solutions) that address critical problems in clinical medicine and population health. Dr. Lenert will present the panel’s second practical PGHD integration example. He will present on apps developed at MUSC for the remote care of patients with COVID-19. Built on MUSC’s strong foundation of telehealth care, Dr. Lenert will describe the development of EHR tools for prescribing either a tethered EHR-portal app or iOS and Android apps for remote care of COVID-19 patients’ symptoms, pulse oximetry data, and temperature as well as alerting telehealth nurses of changes in patient’s status. One of the unique features of this architecture was the use of Clinical Decision Support (CDS) Hooks based triggers within the EHR to prescribe iOS and Android apps. “Deep-linkages” were used to automatically configure iOS and Android apps to work with sensors providers selected in the EHR. The tethered app was used for the care of more than 1300 patients during the “first wave” of the pandemic. Work with the iOS and Android apps was expanded to create a novel system for surge capacity for critical care using an Internet-of-Things approach.

Danielle Lavallee, PharmD, PhD (discussant) is the Scientific Director at the British Columbia Academic Health Science Network. Her research focuses on patient-centered approaches to healthcare delivery including the use of PGHD to support shared decision-making. Dr. Lavallee’s work includes developing guidance for health systems on the integration of PRO data (ePRO toolkit), understanding opportunities, barriers, and priorities for health system integration of PGHD, and evaluating the use of mobile health for supporting post-operative care management. Dr.

60
Lavallee will lead the panel and the audience, in a conversation about challenges and opportunities in collecting and integrating PGHD with EHRs, focusing on technical and organizational considerations. Discussion questions may include:

- What are the pros and cons of different technical approaches presented today?
- What is required in terms of expertise and resources to integrate PGHD with EHRs?
- How might you use the SMART Markers framework?
- What is the utility of PGHD?
- What qualities should PGHD have?
- How could you decide which PGHD to collect and integrate?
- What are some potential impacts of using PGHD on clinic workflow?
- What type of conditions do (and do not) need long-term PGHD collection? For those that do, how do you keep patients engaged in submitting PGHD?
- How to streamline patient authentication?
- What evidence will support continued advancement for integrating PGHD into care?
- How to collect and use PGHD with a health equity lens?

**Topic Rationale:**

Providing person-centered care to ensure that health services are respectful of, and responsive to, the preferences, needs, and value of people is an HHS priority. The increase of social networking, cloud-based platforms, connected devices, and smartphone apps provide simplified means for people to collect and share data outside the traditional clinical environment. While these technologies are promising, the ability for ambulatory care practices to successfully collect these data in collaboration with patients, transfer data to their EHRs, and use them effectively to inform treatment decision making in clinics poses many challenges. The COVID-19 pandemic heightens the need to support ambulatory care practices in this complex process of collecting and using PGHD because many face-to-face visits have been replaced with virtual visits. This panel covers the current state and recommendations of PGHD integration, and features different technical architectures and workflows that support the collection and integration of PGHD.

**Statement of Agreement to Participate**

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

**References:**

Proven Methodologies for Accelerating Adoption of HL7® FHIR®

Charles Jaffe, MD, PhD, FACMI; CEO Health Level 7; Education – Duke University, School of Medicine, M.D., Medicine; Duke University, Ph.D., Experimental Pathology / Computer Science; The Johns Hopkins University, B.A., Chemistry; Affiliations – Visiting Scholar, Division of Biomedical Information, University of California San Diego School of Medicine; Location – Del Mar, California

Steve Kassakian, MD, MS, FACP; Physician Executive, Clinical Informatics at Humana and Assistant Professor at Oregon Health & Science University (OHSU); Education – Oregon Health & Science University, M.S., Biomedical Informatics; Brown University, M.D., Medicine; University of Washington, B.S., Chemistry; University of Washington, B.S., Oceanography; Affiliations – American Medical Informatics Association; American College of Physicians; Location – Portland, Oregon

Heidi Kriz, Assistant Director of Medical Policy and PA Transformation Lead, Cambia Health Solutions/Regence Blue Cross Blue Shield; Education – Barry University, M.P.H., Epidemiology and Healthcare Administration; University of Idaho, B.S., Nutrition Science; Affiliations – Cambia Health Solutions/Regence Blue Cross Blue Shield; Location – Portland, Oregon

Viet Nguyen, M.D., Technical Director, HL7 Da Vinci Project and Founder, Stratametrics; Education – University of Utah School of Medicine, M.D., Medicine; UC San Diego, B.S. Bioengineering; Affiliations – HL7; Location – Salt Lake City, Utah

Anna Taylor, Director of Operations, MultiCare Connected Care; Education – University of Washington, College of Engineering, B.S., Technical Communication (now called Human Centered Design and Engineering); University of Washington, School of Nursing and Medicine, M.S., Clinical Informatics and Patient Centered Technologies; Affiliations – HIMSS, NAACOS, HL7, Baldrige Board of Examiners; Location – Tacoma, Washington

Abstract

The industry panel covers the journey of harnessing a multi-stakeholder collaborative process to form a FHIR Accelerator which leverages HL7® FHIR® standards to transform health care, highlighting change management and the paradigm shift in how organizations work internally and with their business partners.

After this session, the learner should be better able to:

- Learn how a multi-stakeholder initiative, the HL7 Da Vinci Project, is helping the industry use standards to solve health care interoperability problems and meet federal requirements
- Understand how the multi-stakeholder FHIR Accelerator methodology could be adopted by other communities with minimal change.
- Learn how to engage with the Da Vinci Project, join the growing collaborative community and directly access the free and open resources.
- Learn about the challenges, benefits and lessons learned of real-world implementations of the standards-based strategies and how organizations can use Da Vinci work to solve problems and meet federal interoperability and access rules.
General Description

During this 90-minute session, the panel will discuss the formation of the HL7 Da Vinci Project as a FHIR Accelerator. Via several convener and implementer viewpoints, the presenters will detail the journey of assembling and harnessing a multi-stakeholder collaborative process to form an Accelerator to leverage standards. By bringing providers, payers and partners together, the collective expertise and efforts of industry experts are leveraged to identify and solve problems using technology. The panel will share all aspects of the journey, from the initial concept of convening an Accelerator to identifying stakeholders, outlining processes and ultimately launching implementations to meet industry data-exchange needs as well as federal interoperability and patient access rules.

The panel will also share how Symposium attendees can engage with the Da Vinci Project, join the growing collaborative community and directly access the free and open resources. Lessons learned from the project are instructive for other organizations and communities of practice who want to implement FHIR in their own workflows. The journey fundamentally is not about technology but transforming health care, highlighting change management and the paradigm shift in how organizations work internally and with their business partners.

Dr. Viet Nguyen, Technical Director for Da Vinci, will facilitate the panel. He has been with Da Vinci Program Management Office since the inception of the Project and will provide an overarching and longitudinal view of the processes. As the Chief Executive Officer of HL7, Dr. Jaffe will discuss the standards organization’s perspective and the genesis for creating the Accelerator program, highlighting Da Vinci’s four-year journey as one of the first FHIR Accelerators that has produced 10 implementation guides to date. Dr. Jaffe will also discuss the challenges that FHIR accelerators put on the standards process as well as the Accelerator program’s future direction and will illustrate how the multi-stakeholder accelerator methodology could be adopted by other communities with minimal change.

Dr. Kassakian will discuss the experience of implementing FHIR from an academic medical center perspective. Ms. Taylor will discuss the experience of implementing FHIR as a multicenter provider organization. Ms. Lindberg will provide a payer perspective in implementing FHIR and coordinating workflow with provider partners. Participants will each detail their experiences as project members and implementers, representing providers and payers. The panelists will outline their organizational history with FHIR, challenges, benefits, lessons learned regarding participating in the Accelerator and FHIR adoption, current implementations and future plans.

Explanation of Why the Topic of Panel is Timely and Anticipated Audience

As the shift to value-based care continues, the need for payer-provider collaboration has become essential. By bringing providers, payers and partners together to identify and solve problems using standards-based strategies, industry can address pain points and improve care, reduce burden, streamline and automate workflow, liberate data and measure quality. The maturation of HL7 FHIR and the federal interoperability and patient access rules to implement FHIR has provided a catalyst for increased adoption and incentive to derive value from FHIR and interoperability.

The anticipated audience for this panel includes physicians, nurses, dentists, pharmacists, and other clinicians; health information technology professionals; computer scientists and systems developers; policy makers; public health professionals; biomedical engineers and bioinformaticians; consultants and vendor representatives; academic researchers and scientists; and other professionals involved in the collection and dissemination of health information.

Discussion Questions

Is your organization participating in a multi-stakeholder collaborative effort like one of the HL7 Accelerators?

If so, what are the benefits?

What challenges have you experienced and how did you overcome them?

How did you define the scope of your project?

What factors did you consider when selecting your use cases?

How did you agree upon priorities?
Panel Organizer Statement

All panelists named in this submission have agreed to take part on this panel during the AMIA 2021 Annual Symposium, which will occur October 30 – November 3, 2021 in San Diego.
Panel Proposal: Privacy-Preserving Federated Biomedical Data Analysis

Xiaoqian Jiang, PhD, 1 Luca Bonomi, PhD, 2 Jaideep S. Vaidya, PhD, 3 Li Xiong, PhD 4 Lucila Ohno-Machado, MD, PhD, 2
1 School of Biomedical Informatics, University of Texas Health Science Center at Houston, Houston, TX
2 Dept. of Biomedical Informatics, University of California San Diego Health, La Jolla, CA
3 Dept. of Management Science and Information Systems, Rutgers University, Newark, NJ
4 Dept. of Biomedical Informatics, Mathematics, and Computer Science, Emory University, Atlanta, GA

Abstract

Federated learning has emerged as a powerful complement to centralized healthcare data repositories due to their flexibility, nimbleness, and decentralized gatekeeping capabilities. Many biomedical data networks (e.g., pSCANNER1, PCORNet2, eMerge3, SCOR4, etc.) have been established to enable the use of electronic health records (EHR) and genomic data for research. In these settings, data are often distributed across different institutions. Due to privacy or other concerns, patient-level data cannot be transmitted outside some institutions, so a mechanism to support data analysis without sharing patient-level data is needed. This naturally leads to the development of privacy-preserving federated biomedical data analysis. Various technologies have been emerging to support studies across distributed networks. The goal of this panel is to provide the AMIA community with 1) an overview of current progress and technical developments in federated biomedical data analysis; 2) discussions of novel techniques for computing in horizontally or vertically partitioned datasets; and 3) insights into practical deployment, current challenges, and opportunities.

General Description of the Panel and the Issues

Various organizations such as governmental agencies, hospitals, and financial companies collect and share various person-specific data for research and business purposes. Often, data from different sources need to be integrated to gain better insights and deliver high-quality research outcomes. The Shared Pathology Informatics Network (SPIN) initiated by the National Cancer Institute aimed to provide an interface to cancer researchers to access pathology specimens’ data stored across multiple healthcare institutions. The pSCANNER clinical data research network (CDRN) aims at connecting 24 million patient data from geographically distant institutions to facilitate clinical and comparative effectiveness research. Data analysis can be very straightforward if biomedical data of interest are collected and deposited in a centralized repository. Alternatively, institutions can use a federated data analysis framework to obtain information on demand. The challenge here is to simultaneously enforce privacy (protected health information is under the regulation of HIPAA privacy and security rules and the common rule), while maintaining data utility (ideally, this translates into producing the same results as if data analyses were conducted in a centralized repository). These challenges call for the development of advanced health information technologies to support federated data analysis. In this proposed panel, we will focus on the discussion of federated data analysis and the protection of its research outcomes.

Structure of the Proposed Panel

The goal of this panel is to review recent research in the field of privacy-preserving federated biomedical data analysis. The panel will be hosted by Dr. Lucila Ohno-Machado, who is the Professor of Medicine and Chair of the Department of Biomedical Informatics at UCSD. Since 2010, Dr. Ohno-Machado has led the NIH-funded integrating Data for analysis, Anonymization and SHaring (iDASH) project. Dr. Ohno-Machado is also the PI of the PCORI-funded pSCANNER clinical data research network, which is designed to allow distributed analyses on data derived from electronic health records of over ten health systems. Dr. Ohno-Machado will provide a review of the needs for federated data analysis in biomedical research and an overview of federated data analysis methodology.
The panel will include four recognized researchers in biomedical data privacy and security. Each of these researchers will provide insight into a different aspect of federated biomedical data analysis and convey their experience with different applications.

The first panelist will be Dr. Xiaqian Jiang from the University of Texas Health Science Center at Houston. Dr. Jiang will share his experience on federated data analysis modeling\textsuperscript{10–13} for the iDASH and pSCANNER projects. The second panelist will be Dr. Luca Bonomi from UCSD. He has been working on privacy-preserving data linkage\textsuperscript{17} and genomic data sharing\textsuperscript{17,18}. He will introduce the applications of these algorithms. The third panelist will be Dr. Jaideep Vaidya from Rutgers University. He will share his expertise in privacy-preserving data mining\textsuperscript{14} and the development of various federated data analysis models for vertically and horizontally partitioned databases\textsuperscript{15,16}. The fourth panelist will be Dr. Li Xiong from Emory University. She was the PI of the PCORI-funded project Building Data Registries with Privacy and Confidentiality for Patient Centered Outcomes Research and will introduce the opportunities and challenges of developing differential privacy technologies\textsuperscript{19} in protecting outcomes of biomedical research in a distributed setting. This is important because the outputs of federated data analysis may still pose privacy risks if disclosed to the public. Thus, the final outcomes (e.g., estimated parameters) also need protection. If we want to expose these outcomes to the public, strong privacy protection like the one provided by differential privacy techniques is necessary to mitigate the risk. Dr. Lucila Ohno-Machado will moderate the panel from UCSD, who is the PI of pSCANNER, iDASH, and RADx-rad projects.

Urgency of the Topic and Relevance for AMIA

Health systems and other institutions that want to participate in research but are unable to transmit patient-level data beyond their local environments or ‘enclaves’ can do so by employing distributed analytics such as the ones we will describe. Patient privacy is protected when only privacy enhanced results of computation need to be transmitted. Using data from multiple institutions has many advantages for research, including increased diversity of populations and increased sample size. For example, when dealing with hypothesis testing and regression models, one may require many samples to increase the statistical power. Oftentimes, data are horizontally distributed (e.g., different institutions need to study similar patients to determine the effectiveness of a new procedure)\textsuperscript{20}. Data owners would like to collaborate for federated data analysis without sharing patient-level data. One way is to use meta-analysis to combine results from local studies through weighted average, but it is often criticized for example, several small studies do not predict the result of a single large study\textsuperscript{21}, and sources of bias are not controlled by the method\textsuperscript{22}. Unlike meta-analysis, federated data analysis guarantees could achieve accurate results as if data analyses are conducted in a centralized manner. In addition to horizontally partitioned data, different stakeholders might have different pieces of patient data, for example, EHR data in clinics, genome data from the same patient population in a sequencing center, claims data in an insurance company. In this setting, data are vertically distributed. Federated data analysis over vertically distributed patient information supports the use of these data in biomedical research, but it is a harder task than computing on horizontally partitioned data.

This panel will address biomedical data privacy for clinical data research networks, where a large portion of the AMIA community is currently involved. The anticipated audiences include but are not limited to healthcare privacy policymakers, biomedical researchers, as well as other stakeholders (e.g., patients, clinicians, etc.)

Discussion questions

Potential discussion questions include, but not limited to:

What are the challenges and opportunities of federated data analysis in protecting the privacy of biomedical research studies?

How to enforce data quality in a federated data analysis?

How to link patient data across vertically partitioned datasets?

What kind of data analyses can be supported in a federated manner?
How can secure computing be used to enhance privacy protection in a federated network?
How can differential privacy techniques be used to support federated data analytics?

Statement from the Organizer
All panelists have approved this proposal’s writing and have agreed to take part in this panel.

References
Evolving Challenges in Patient Matching

Abel N. Kho, MD, MS\(^1\), Shaun J. Grannis, MD, MS, FAAFP\(^2\), Adam Culbertson, MS, MS\(^1\), Molly Murray, MS\(^3\), Yu Deng\(^4\), BS

\(^1\)Center for Health Information Partnerships, Northwestern University, Chicago, IL, USA; \(^2\)Regenstrief Institute, Indianapolis, IN, USA; \(^3\)Pew Charitable Trusts, Philadelphia, PA, USA

Abstract

COVID-19 data tracking and reporting efforts have highlighted the vital importance and ongoing challenges associated with linking patient data using demographic variables. Currently, match rates between organizations can be as low as fifty percent, resulting in barriers in information exchange, introducing safety risks, and increasing costs. The difficulty of patient matching is multifactorial and includes the variation in quality and availability of matching variables across healthcare organizations and time as well as restrictions in access to patient unique identifiers due to data security and privacy concerns. Participants in this panel discussion will present 1) current lessons learned and evolving challenges to successful patient matching 2) current state-of-art patient matching approaches and promising new methods 3) recent policies and legislation affecting patient matching 4) privacy, security, and ethics issues over patient matching especially during COVID vaccine tracking 5) the changing nature of patient matching features available over the past 10 years across a diversity of care settings and how these features were impacted by the COVID-19 pandemic.

Introduction

The Health Information Technology for Economic and Clinical Health (HITECH) Act of 2009 and the Federal Health IT Strategic Plan set the adoption and meaningful use of electronic health records (EHR) as the key objectives (1). Since 2009, the use of basic EHR systems and the possession of certified EHR technology has increased significantly (2). Despite the wide adoption of EHR, the lack of interoperability between different healthcare organizations has hampered care coordination, health information exchange and more efficient patient care (3). Such challenges have compounded during the COVID-19 pandemic where fast data tracking, reporting efforts are required. Accurate patient matching is the key to improve EHR interoperability. The Office of the National Coordinator (ONC) for Health Information Technology reported that the rates of record matching can be as low as 50 percent among provider records exchange (4).

The challenge of accurately matching patient records comes from many aspects. Concerns over data security and more restricted government policy on Protected Health Information (PHI) limits the access to patient unique identifiers that many linkage methods rely on (5). Variation in data quality also imposes a significant challenge to patient matching (6,7). Ideally, linkage variables should be unique, accurate, complete (non-missingness), and consistent across sites and time. However, in real world data, erroneous entry, missingness, and information updates (e.g. patient address) are common for many demographic variables. Data standardization is another common problem in patient matching (8). Matching variables are often captured in varying formats in healthcare organizations and health information systems (7). Currently, several organizations including the Agency for Healthcare Research and Quality (AHRQ) and the ONC, have proposed best practice approaches for data standardization (8). However, there is no consensus on which specific data standardization method should be used.

In this session, the panelists will provide an overview of the evolving field of patient matching, including current and emerging trends in patient matching methods (e.g. unique patient identifiers, empowered patient solution, data standardization, referential matching), recent policies and legislation related to patient matching, potential security and privacy risks that inconsistent patient matching may cause, as well as recent work quantifying the changing nature of patient features available over the past 10 years across a diversity of care settings and how these features were impacted by the COVID-19 pandemic.

Panelists and Topics

1) Moderator: Abel Kho, MD, MS, FACMI is a Professor of Medicine and Director of the Center for Health Information Partnerships and the Institute for Augmented Intelligence in Medicine at Northwestern University. Dr. Kho has expertise in integrating EHR data across diverse care settings, with a focus on privacy preserving record linkage methods, as part of several national networks including PCORnet 2.0, the All of Us Research Program, and N3C. He will moderate the discussion and provide an overview of the changing nature of person identity and some of the persisting challenges to successful patient matching.
2) Shaun J Grannis, MD, MS, FACMI is the Sam Regenstrief Professor of Medical Informatics, Indiana University School of Medicine, and Vice President for Data and Analytics, Regenstrief Institute. Dr. Grannis collaborates with national and international healthcare stakeholders to advance interoperability and data sharing capabilities. He has provided identity management consultancy to low- and middle-income countries, as well as the World Health Organization and the Office of the National Coordinator for Health Information Technology. Dr. Grannis will present recent methods for identifying optimal fields for standardization and matching. He will describe how these methods affect real-world patient matching accuracy using datasets from various data sources, including public health data sources (including COVID-19 laboratory results and vaccinations), health information exchange, vital records, and newborn screening data.

3) Adam Culbertson, MS, MS, has worked on patient matching for several years. He served a term as the HIMSS Innovator in Residence working with the Office of the National Coordinator for Health IT and the CTOs office to advance patient matching through the patient matching challenge. He was the lead author on the Building Blocks of Interoperability, a study of the different data elements that can be used for patient matching. He has contributed to numerous articles, policy forums, and a frequent speaker on the topic patient matching.

4) Molly Murray, MPA, is an Officer on Pew’s Health Information Technology project. Her work focuses on research and advocacy related to interoperability and patient matching. She will provide an overview of current regulations and policies, as well as recent proposed legislation relevant to patient matching.

5) Yu Deng, BS, is a PhD candidate in biomedical informatics at Northwestern University. Her research focuses on improving patient outcome and quality of care using EHR data. Yu Deng will present her recent work showing how the COVID-19 pandemic has affected the types of patient features currently collected.

Panel Rationale

Currently the ability to match a patient's records across the medical system is foundational to improving care and lowering cost. The issue of patient matching has been identified as one of the leading barriers to interoperability in the Nations Road Map to Health IT (3). In 2020 Senators Maggie Hassan (D) and Bill Cassidy(R), introduced a bipartisan bill called the Patient Matching Improvement Act to help make improvements to patient matching by requiring the use of the “United States Postal Service (USPS) standard, to improve data quality of address features used for matching (9). The current COVID pandemic has accelerated the need to solve the nation's patient matching bottleneck and at the same time brought privacy and security issues to the forefront. The need to contract trace and maintain immunization registries for public health, and track who has received immunizations (first and second dose) while maintaining data privacy and security have made the issue a crucial challenge in our COVID response. There is growing consensus that our nation's patient matching challenge must be solved. However, there still is no consensus on what approach would best link the disparate patient data that is scattered across the healthcare system. This panel discussion will provide a forum for leading thought leaders in industry, academia and government to convene to advance the national patient matching dialogue.

This panel will explore the evolving challenges and opportunities to the art of patient matching; the impact of data standardization on patient matching; the variation in availability of patient features over the past 10 years and across multiple healthcare organizations, the challenges and opportunities of patient matching raised by the COVID-19 pandemic, as well as the ethics and privacy issues with patient matching.

This panel comprises a multi-disciplinary team reflecting complementary perspectives on the policy and technology required to develop effective informatics approaches to improve patient matching. Dr. Kho will moderate the session and provide an overview of the current and persisting challenges in patient matching. Dr. Shaun Grannis will compare and contrast the impact of data standardization vs algorithm performance on patient matching. Mr. Culbertson will describe the changing nature and availability of patient features for patient matching from a diversity of care settings. Yu Deng will present recent work showing how the COVID-19 pandemic has affected the types of patient features currently collected. Molly Murray will provide an overview of current policies and recent proposed legislation relevant to patient matching.

The anticipated audience for this presentation are individuals interested in understanding the evolving challenges to patient matching, the current trends in patient matching algorithms, the impact of data standardization on patient matching, and the dynamic shifts in demographic feature availability over the past 10 years and the recent impact of COVID-19 on these features.
Discussion questions

The goal of the panel will be to present expert and research informed viewpoints on specific aspects of record linkage and engage the audience in subsequent discussion. Each speaker will present for 15 minutes. Then once the short presentations are completed there will be an interactive panel with audience participation and questions and comments from the audience. Topic and questions for discussion will include:

- What are the emerging technologies and algorithms that can be leveraged for the purpose of improving the availability and quality of patient features available for matching?
- How can patient matching improve COVID vaccine tracking? How does HIPAA, ethics and other concerns come into play in COVID vaccine tracking?
- How does the availability of demographic variables for matching change in the past 10 years and what is its impact on patient matching?
- How can data standardization algorithms (e.g. the USPS standard) be applied to improve the quality of features used for patient matching?
- What will the latest innovations be in patient matching? What are some alternatives (e.g. the CMS Blue Button initiative, patient unique identifier) in patient matching in addition to data quality improvement and data standardization?
- How can we better engage patients in the patient’s matching process to ensure patients data is their own and how they can protect their privacy through efficient consent policies and technology?
- What data privacy, data security and ethics issues do different patient matching methods bring?
- What are the policy considerations and the role of federal agencies (e.g. ONC, NIST, and CMS) to improve patient matching?

Organizer Statement

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

References

The COVID-19 Pandemic as Catalyst: The Acceleration of Registry Science in the 21st Century to Address Novel Challenges

Steven E. Labkoff, MD, FACP, FACMI, FAMIA1, Leon Rozenblit, JD, PhD2, Rachel Richesson, Ph.D., MPH, FACMI3, Helen Burstin, MD, MPH, MACP

1. The Multiple Myeloma Research Foundation, Norwalk, CT, 2 Prometheus Research, an IQVIA Company, New Haven, CT, 3 University of Michigan, Medical School, Ann Arbor, MI, 4 Council of Medical Specialty Societies, Chicago, IL

Abstract:
With the rise of the pandemic, there has been a renewed focus on the construction and utilization of patient registries. Having evolved gradually for over 75 years, registry use has been reshaped by this crisis. Registry use continues to grow because EHRs, while important for patient care and administrative issues to run a healthcare institution, often do not collect enough of the right data about specific medical issues. Clinical registries have been stood up and repurposed to address the need for rapid information on the appropriate management of COVID-19, especially for patients with secondary conditions, including cancer and autoimmune diseases. For example, in rare diseases, registries are often the only mechanism to collect enough relevant data to estimate the prevalence, understand the presentation and progression. Registry Science is the branch of Informatics that bridges the gap between clinical care and clinical research through high-quality, focused data assets. This panel will discuss how registry science has emerged and advanced in the past 20 years and highlight important use cases and informatics and data challenges, up to and including how the COVID-19 pandemic is raising the profile and need of Registry Science.

Introduction:
As the US has seen a meteoric rise in the use of EHRs over the last 15 years, there has been a concomitant rise in the construction and utilization of medical registries. One might think that this would not be the case as EHRs are generally considered to be the source of all clinically-relevant medical data for the care of patients. However, as it turns out, in 2021, EHRs are not nearly as good at generating data sets for clinical research as are registries. The rise of personalized medicine, the need for research into rare diseases, FDA REMS mandates for safety, and above all, the COVID-19 pandemic has all contributed to the growth and need for registries.

Where EHRs collect data for clinical care, they seldom collect as complete or focused a data set as is needed to answer research questions about a specific disease or cohort. One problem that registries need to solve is to identify and aggregate patients with the same diagnosis and bring together relevant and detailed data for analysis. EHRs, while excellent at collecting data for analysis and clinical care of the individual patient, often fall short, especially for a rare disease or rare event cohorts. For a rare disease that only has an incidence of 30,000 new cases, as is the case with multiple myeloma, no single institution likely has enough cases to perform sufficient clinical research. Likewise, in the case of a massive epidemiologic study, aggregating data from “all-comers” from all EHRs remains a challenging, if not
impossible task. Further, because EHRs are designed for documenting the care of an individual patient, they don’t routinely assess or record information that may be critical for research in a specific area.

In this panel, panelists with extensive and varied experience will describe the evolving paradigm for clinical registries, including ongoing challenges related to data acquisition and standardization, harmonization, cloud-based computing, and the growing importance of patient-generated data, including patient-reported outcomes. Panelists will discuss how the nature and needs for registries have evolved through COVID-19 and will consider their roles in understanding and treating rare diseases, detecting adverse drug events, and delivering personalized medicine. They will also discuss how modern registries collect and aggregate data from disparate sources as well as the data analytics needed to deal with large genomic and immunome files that are becoming standard in these registries.

Panelists:
Steven Labkoff, MD, FACP, FAMCI, FAMIA is the Chief Data Officer of the Multiple Myeloma Research Foundation in Norwalk, Connecticut. He is the program sponsor and chief architect of the Multiple Myeloma Research Foundation’s (MMRF) CureCloud Direct-to-Patient Registry (CC-DTP). He is responsible for the MMRF’s 2 petabytes of myeloma data. Previously he was Head of Medical Strategy at Purdue Pharma where he managed Medical Affair’s Big Data Program. Previously he was VP Life Sciences at Intelligent Medical Objects where he worked on vocabulary and ontology tools for Life Sciences. As AstraZeneca’s Executive Director of Research and Development Informatics he built 3 informatics departments including Real World Evidence, Biomarkers, and Clinical Trial Decision Support. He spent 13 years at Pfizer in Business Technology and Medical Affairs where he built the informatics infrastructure for the Infectious Diseases Institute, Kampala, Uganda, an outpatient HIV hospital. He is an ACMi fellow and has served on AMIA’s Executive Committee and served on various AMIA working groups and committees. He founded the AMIA Industry Advisory Counsel and ran both national and international workstreams on Secondary Healthcare Data Use.

Leon Rozenblit is Senior Director (and founder) of Prometheus Research, an IQVIA company, a registry-focused informatics company in New Haven, Connecticut, and he has been directly involved in a leading role in designing and building dozens of registries across the US. On a day-to-day basis, his organization builds and manages patient registries and agile, integrated data hubs for multiple medical specialty societies and patient advocacy groups, and is the technology partner on the CureCloud Direct-to-Patient registry. Having to take into account varying data standards and data models is a daily concern for his organization and his clients. He will speak to the myriad of issues that comprise the successful implementation of such repositories, from a standard and data model perspective.

Rachel Richesson, Ph.D., MPH, FACMI, is a professor of Informatics and Learning Health Sciences at the University of Michigan Medical School. She has coordinated the implementation of data standards for a number of multi-national multi-site clinical research and epidemiological studies, including the NIH Rare Diseases Clinical Research Network (RDCRN), Type 1 Diabetes TrialNet, and The Environmental Determinants of Diabetes in the Young (TEDDY) study, and the national distributed Patient-Centered Outcomes Research Network (PCORnet). Dr. Richesson helped design the RDCRN Contact Registry to facilitate research in over 150 different rare diseases, and she serves on advisory boards for several rare disease registries and patient advocacy groups. Currently, Dr. Richesson co-leads the Electronic Health Records Core
for the NIH Health Systems Research Collaboratory, which is developing best practices for using EHR data in pragmatic clinical trials, including data quality assessment, clinical phenotyping methods, and selection of study endpoints. She is highly engaged in national discussions related to the development of Health Information Technology standards and policy to support clinical research, and participates in multiple standards development organizations, including Health Level Seven (HL7) Patient Empowerment and Vocabulary working groups.

Helen Burstin, MD, MPH, MACP is the Chief Executive Officer of the Council of Medical Specialty Societies (CMSS) which represents 45-member specialty societies with collective membership of 800,000 U.S. physician members. CMSS works to support and strengthen specialty societies and catalyzes improvement through convening, collaboration, collective voice, and action. Many CMSS member societies support clinical registries. CMSS has identified the digital transformation of registries and research as a major strategic priority. With support from the Gordon and Betty Moore Foundation, CMSS recently hosted a webinar series on COVID-19 and Clinical Registries. Dr. Burstin formerly served as Chief Scientific Officer of The National Quality Forum (NQF). A former AMIA board member, she now serves on the boards of AcademyHealth and the Society to Improve Diagnosis in Medicine. Dr. Burstin is a Clinical Professor of Medicine at George Washington University.
Panel Proposal: Multi-Modal Data Science for Healthcare: State of the Art, Challenges, and Opportunities

Yuan Luo, PhD\textsuperscript{1}, Fei Wang, PhD\textsuperscript{2}, Benjamin Glicksberg, PhD\textsuperscript{3}, Jessilyn Dunn, PhD\textsuperscript{4}, Nigam Shah, MBBS PhD\textsuperscript{5}

\textsuperscript{1}Moderator, Northwestern University, Chicago, IL, USA
\textsuperscript{2}Cornell University, New York, NY, USA
\textsuperscript{3}Icahn School of Medicine at Mount Sinai, New York, NY, USA
\textsuperscript{4}Duke University, Durham, NC, USA
\textsuperscript{5}Stanford University, Palo Alto, CA, USA

Abstract

The United States Precision Medicine Initiative aims to utilize information from multiple modalities—phenotypic, omic, imaging, and environmental—to develop an individualized and comprehensive view of a patient’s pathophysiologic progression, to identify unique subtypes of patients, and to administer personalized therapies. The rapid growth of multiple data modalities, when linked to the right patients, may provide a comprehensive view of the patients’ pathophysiology and offer a basis for meaningful subtyping of these patients. This panel will draw on their extensive experiences working with multi-modal data for healthcare problems to provide their perspectives on the best practices and pitfalls to avoid.

Background

Precision medicine aims to utilize information from multiple modalities (e.g., phenotype and omics) to develop an individualized and comprehensive view of a patient, to identify unique disease endotypes (sub-phenotypic patient group sharing pathophysiologic and mechanistic processes of a disease), and to administer personalized therapies\textsuperscript{1}. Existing efforts are often based on only a pre-selected set of biomarkers from a single modality. However, the different data modalities, when linked to the right patients, may provide a holistic view of the underlying pathophysiology of these diseases and offers a basis for meaningful subtyping. The rich information contained in multi-modal patient data sources asks for effective and efficient computational approaches for analyzing them.

To effectively integrate multiple data modalities, significant challenges arise from each modality separately and from the integration process. Regarding the phenotype data, the higher-order temporal trend features need to be properly modeled instead of modeling of atomic features (snapshot measurements). Although recent deep learning methods (e.g., recurrent neural networks) are increasingly used for modeling data longitudinality, these systems are often regarded as black-boxes and lack interpretability. Moreover, most data-driven methods are not able to capture a group of abnormal physiologic trends that can reflect (multi-)organ dysfunctions.

Regarding omics, GWAS studies have identified thousands of SNPs (atomic features) associated with different diseases or phenotypes. However, little is known about the molecular mechanisms of how these variants contribute to disease pathogenesis. One of the major roadblocks is that the majority of GWAS SNPs are located in the non-protein coding regions, most of whose functions remain poorly understood. For example, in hypertension, while GWAS from numerous consortia identified 118 independent loci associated with hypertension and additional loci associated with heart failure; exome sequencing, proteomic, and transcriptomic analyses of heart failure and hypertension are less mature and less informative on molecular mechanisms. Linking genomic and transcriptomic signatures to disease endotypes may provide additional mechanistic understanding of pathogenesis and identify future targets for therapy.

Intended Audience

Biomedical informatic researchers and practitioners.
Topics for Discussion

Yuan Luo, PhD, will talk about his experiences in integrating multi-modal healthcare data when investigating different diseases. For example, he will talk about using hybrid matrix factorization to integrate both the phenotypic and genomic features to derive informative subgrouping of hypertension patients. He will also talk about their efforts in using autism spectrum disorder (ASD) to demonstrate the clinical utility of integrating omic and phenotypic data for endotype discovery. By combining healthcare claims, electronic health records, familial whole-exome sequences, and neurodevelopmental gene expression patterns, one can identify a subgroup of patients with dyslipidemia-associated ASD, which informs early screening and early intervention.

Benjamin Glicksberg, PhD: There have been many efforts to try to apply machine learning to aid in the COVID-19 pandemic. In this presentation, I explore several machine learning studies conducted to predict clinically-relevant events for COVID-19 patients at the Mount Sinai Health System – one of the largest and most diverse health systems in New York City. I identify strategies and novel frameworks to improve predictive performance of these models, which could be applied outside of the context of COVID-19. I describe how these research efforts utilized multi-modal patient data (Electronic Health Records, free-text notes, biomedical imaging, and electrocardiogram waveforms) and were applied across several clinical domains.

Jessilyn Dunn, PhD: Mobile and Digital Health are poised to not only make it possible to rapidly detect, prevent, and manage many diseases, but also ensure equitable healthcare access. However, progress has suffered because mobile and digital health data presents unique bioinformatic challenges; Specialized tools are needed for data annotation, pre-processing, organizing, and exploratory and hypothesis-driven analysis and validation that will enable machine learning and multi-modal data integration. Our team built the Digital Biomarker Discovery Pipeline (DBDP.org), which transforms large volumes of mHealth data into digital biomarkers for disease detection, monitoring, and prevention. Digital biomarkers are digitally collected data (e.g. mean resting heart rate) that are transformed into indicators of health outcomes (e.g. fitness) and can be used to provide biomedical insights or improve health decision-making.

Nigam Shah, MBBS PhD: Widespread adoption of electronic health records (EHRs) has fueled the development of using machine learning to build prediction models for various clinical outcomes. Patient representation schemes inspired from language modeling increase the accuracy of clinical prediction models by transferring information learned from the entire patient population to the task of training a specific model. However, as the development of predictive models spans multiple modalities, it is essential to investigate advanced representation schemes that can handle EHR as well as imaging data together. Using the example of predicting long-term outcomes after pulmonary embolism, we will discuss the trade-offs required in learning models spanning both imaging and EHR data. We will discuss the key considerations for evaluating the usefulness of patient representations spanning multiple modalities.

Fei Wang, PhD, will talk about deep clinical risk prediction with patient data from multiple modalities, and the siamese deep modal architecture for the scenarios where the sample sizes are not large.

Timing for the Panel

Multi-modal data science has shown great successes in many domains with big data sets. However, its impact on healthcare applications is still very limited. As diverse large biomedical data are being collected, people can foresee potential success of combining them and triangulating the various streams of evidence in healthcare. We assemble a group of pioneers in researching advanced machine learning models to combine multi-modal data for healthcare problems to provide their perspectives in this topic and collectively discuss the best practices and pitfalls to avoid multi-modal data science for addressing healthcare problems.

Discussion Questions

• What are the major successes of multi-modal data science in healthcare?
• What are the challenges in using machine learning models for combining multi-modal biomedical data sources?
• How to provide interpretable models in the presence of multiple modalities of data?
• What are the common mistakes in using multi-modal data for addressing healthcare problems?
• What are practical solutions of data missingness as this problem is amplified with the integration of several modalities as described?

• In addition to interpretability, what are the challenges with multimodal representations for validation and applying models to new patients?

• How do differences in data veracity across modalities affect implementation and performance of multi-modal models?

Panel Lineup

Moderator: Yuan Luo, PhD, is an Associate Professor at Department of Preventive Medicine, Division of Health Biomedical Informatics at Feinberg School of Medicine in Northwestern University. He is Chief AI Officer at Northwestern University Clinical and Translational Sciences Institute and Institute for Augmented Intelligence in Medicine. His research interests include machine learning, natural language processing, time series analysis, computational phenotyping and integrative genomics, with a focus on biomedical applications. He won the American Medical Informatics Association (AMIA) New Investigator Award in 2020. He is currently an editor with JAMIA Open, JBI, Plos One, JHIR. He served on AMIA Membership and Outreach Committee. Dr. Luo is a fellow of the American Medical Informatics Association (AMIA).

Fei Wang, PhD, is currently an Associate Professor of Health Informatics in Department of Population Health Sciences. His major research interest is developing AI algorithms for solving medical problems, such as clinical risk prediction and disease subtyping. So far he has published more than 250 papers on top venues in AI (such as AAAI, KDD), clinical medicine (such as JAMA Internal Medicine and Annals of Internal Medicine) and health informatics (such as AMIA and JAMIA). His papers have received around 15,000 citations and his H-index is 60. He has received numerous awards including NSF CAREER award, IEEE health informatics leadership award, as well as the awards from industries including Google faculty research award, Amazon machine learning for research award (twice) and Sanofi iDEA award. He has been the PI on numerous grants from federal agencies NSF, NIH, ONR and private foundations such as MJFF. Dr. Wang is a fellow of the American Medical Informatics Association (AMIA).

Benjamin Glicksberg, PhD, Benjamin Glicksberg is an Assistant Professor of Genetics and Genomic Sciences and a member of the Hasso Plattner Institute for Digital Health at the Icahn School of Medicine at Mount Sinai. He helped found and serves as the head of data science for the Mount Sinai Clinical Intelligence Center. He uses machine learning on bio- and clinical informatics frameworks to personalize medicine. This work often involves integrating multi-omic data including genomics, imaging, and clinical information, particularly involving Electronic Health Records. He completed his PhD in Neuroscience at the Icahn School of Medicine at Mount Sinai in 2017 and post-doctoral work at the University of California, San Francisco in 2019.

Jessilyn Dunn, PhD, Jessilyn Dunn is an Assistant Professor of Biomedical Engineering and Biostatistics Bioinformatics at Duke University. Her primary areas of research focus on biomedical data science and mobile health; her work includes multi-omics, wearable sensor, and electronic health records integration and digital biomarker discovery. Dr. Dunn is the Director of the BIG IDEAs Laboratory, whose goal is to detect, treat, and prevent chronic and acute diseases through digital health innovation. Dr. Dunn was an NIH Big Data to Knowledge (BD2K) Postdoctoral Fellow at Stanford and an NSF Graduate Research Fellow at Georgia Tech and Emory, as well as a visiting scholar at the US Centers for Disease Control and Prevention and the National Cardiovascular Research Institute in Madrid, Spain.

Nigam Shah, MBBS PhD Nigam Shah is Professor of Medicine (Biomedical Informatics) at Stanford University, and serves as the Associate CIO for Data Science for Stanford Health Care. Dr. Shah’s research focuses on combining machine learning and prior knowledge in medical ontologies to enable the learning health system. Dr. Shah was elected into the American College of Medical Informatics (ACMI) in 2015 and was inducted into the American Society for Clinical Investigation (ASCI) in 2016. He holds an MBBS from Baroda Medical College, India, a PhD from Penn State University and completed postdoctoral training at Stanford University.

Panel Organizer Statement: We affirm that all panel participants have agreed to participate and have contributed to the preparation of this document.
Panel Proposal: Graph Based Machine Learning for Healthcare: State of the Art, Challenges, and Opportunities

Yuan Luo, PhD\textsuperscript{1}, Fei Wang, PhD\textsuperscript{2}, Marinka Zitnik, PhD\textsuperscript{3}, Shuiwang Ji, PhD\textsuperscript{4}
\textsuperscript{1}Moderator, Northwestern University, Chicago, IL, USA
\textsuperscript{2}Cornell University, New York, NY, USA
\textsuperscript{3}Harvard University, Boston, MA, USA
\textsuperscript{4}Texas A&M University, College Station, TX, USA

Abstract

Analysis with graphs is a classic topic in data mining and many techniques have been proposed in the past. Recently, because of the rapid development of data mining and knowledge discovery, many novel graph based machine learning algorithms have been proposed and successfully applied in healthcare. This panel will reflect on the graph based machine learning algorithms developed recently and how they have been applied in healthcare. They will discuss on using the graph based machine learning models to integrate knowledge graphs with real world data. The panel will also discuss a set of potential issues and challenges such as interpretability, fairness and security.

Background

Graph is a natural representation encoding both the features of the data samples and relationships among them. For example, biomedical relations (including events) can be universally represented as graphs by converting biomedical concepts to nodes and syntactic/semantic links to edges. Other propositional representations may require specific interpretation of the graphs. For instance, representing the negation of a proposition may require the introduction of nested graphs, and to give special semantics to a relation labeled NOT. Furthermore, although composition leads to complexity (e.g. n-ary relations or nested relations), by adopting a graph-based representation, we can focus on common graphical patterns that provide good ways to capture relations. Graph based machine learning has been a popular tool for analyzing interconnected data entities. Analysis with graphs is a classic topic in data mining and many techniques have been proposed in the past. In recent years, because of the rapid development of data mining and knowledge discovery, many novel graph based machine learning algorithms have been proposed and successfully applied in a variety of areas. Different from conventional statistical methodologies which typically assumes the data objects are independently and identically distributed (i.i.d.), graph algorithms can model both the data features and relationships simultaneously. This is particularly important for applications with complex and heterogeneous data.

The goal of this panel is to reflect on the graph based machine learning algorithms developed recently and how they have been applied in healthcare. In particular, the panel will cover both the technical advances and the application in healthcare. On the technical aspect, we will discuss deep network embedding techniques, graph neural networks, knowledge graph construction and inference, graph generative models and graph neural ordinary differential equation models. On the healthcare side, we will discuss how these methods can be applied in predictive modeling of clinical risks (e.g., chronic disease onset, in-hospital mortality, condition exacerbation, etc.) and disease subtyping with multi-modal patient data (e.g., electronic health records, medical image and multi-omics), knowledge discovery from biomedical literature and integration with data-driven models, as well as pharmaceutical research and development (e.g., de-novo chemical compound design and optimization, patient similarity for clinical trial recruitment and pharmacovigilance). The panel will also discuss a set of potential issues and challenges such as interpretability, fairness and security. In particular, the panel will discuss methods for learning disentangled graph representations that capture rich notions of graph topology. They will highlight advances in few-shot learning, yielding models that generalize to completely new graphs and never-before-seen labels using only a handful of nodes or edges. We will ground the graph based ML in examples (e.g., genetic medicine) that have helped healthcare to spark the interest in a wide audience. We will also ensure there is balance between technical details, application and security.

Intended Audience

Biomedical informatic researchers and practitioners.
Learning objectives

The panels aims to benefit participants for:

- Understanding graph based machine learning as an effective means to model and predict complex relations in order to advance biomedical science and improve human health (basic level)
- Understanding recent advances in algorithms and methods used in graph based machine learning (intermediate level)
- Addressing the fundamental challenges in graph based machine learning such as producing fair and stable representations that can be interpreted meaningfully (advanced level)

Topics for Discussion

Yuan Luo, PhD, will discuss graph neural network-based method for text classification covering clinical notes and biomedical literatures, including both long text such as documents and short text such as relations in sentences. He will discuss how to automatically extract relational information out of text into a graph representation and how to learn features from graphs. He will demonstrate that these methods lead to progressively improved performance by integrating lexical, syntactic and semantic information.

Fei Wang, PhD, will discuss the potential of machine learning algorithms on deriving hypothesis and improving quality of biomedical knowledge graphs. He will also talk about how to interact between the inference on knowledge graphs and real world data to achieve mutual performance boosting.

Marinka Zitnik, PhD, will discuss recent efforts to expand the scope and ease the applicability of graph representation learning in healthcare. She will discuss methods for learning disentangled graph representations that capture rich notions of graph topology. She will highlight advances in few-shot learning, yielding models that generalize to completely new graphs and never-before-seen labels using only a handful of nodes or edges. She will also mention efforts at the interface of trustworthy AI and graph machine learning that produce fair and stable representations that can be interpreted meaningfully. Finally, she will discuss applications in disease diagnosis, treatment recommendations, and the development of safe and effective therapeutics.

Shuiwang Ji, PhD, will discuss methodological developments of graph neural networks, including expainability, 3D graphs, generation, and applications into molecular property prediction and drug discovery, molecular graph generation.

Timing for the Panel

Graph based machine learning has shown great successes in many domains with big data sets. This panel will summarize the popular graph based mining technologies that have already been widely applied in healthcare research. As diverse large biomedical data are being collected, people can foresee potential success of combining them and triangulating the evidences in healthcare. We assemble a group of pioneers in researching advanced graph based machine learning for healthcare problems to provide their perspectives in this topic and collectively discuss the best practices and pitfalls to avoid in graph based machine learning for addressing healthcare problems.

Discussion Questions

- What are the major successes of graph based machine learning in healthcare?
- What are the challenges in using graph based machine learning models?
- How to provide interpretable models in the presence of graph based machine learning?
- How to build trustworthy graph machine learning models that produce fair and stable representations?
- How to use graph based machine learning models to leverage knowledge graphs to efficiently inference on real world data?
• What are the common mistakes in using graph based machine learning for addressing healthcare problems?

Panel Lineup

Moderator: Yuan Luo, PhD, is an Associate Professor at Department of Preventive Medicine, Division of Health Biomedical Informatics at Feinberg School of Medicine in Northwestern University. He is Chief AI Officer at Northwestern University Clinical and Translational Sciences Institute and Institute for Augmented Intelligence in Medicine. His research interests include machine learning, natural language processing, time series analysis, computational phenotyping and integrative genomics, with a focus on biomedical applications. He won the American Medical Informatics Association (AMIA) New Investigator Award in 2020. His PhD Thesis was awarded the inaugural Doctoral Dissertation Award Honorable Mention by AMIA in 2017. He is currently an editor with JAMIA Open, JBI, Plos One, JHIR. He served on AMIA Membership and Outreach Committee. His publications appear in leading journals including Nature Medicine, JAMIA, JBI, JAMA Open etc. He has published in and served as PC members for top AI and informatics conferences including AAAI, IJCAI, AMIA, AMIA Joint Summits, IEEE ICHI (track chair) etc. His research is funded by multiple NIH R01 and R21 grants (PI or MPI). Dr. Luo is a fellow of the American Medical Informatics Association (AMIA).

Fei Wang, PhD, is currently an Associate Professor of Health Informatics in Department of Population Health Sciences. His major research interest is developing AI algorithms for solving medical problems, such as clinical risk prediction and disease subtyping. So far he has published more than 250 papers on top venues in AI (such as AAAI, KDD), clinical medicine (such as JAMA Internal Medicine and Annals of Internal Medicine) and health informatics (such as AMIA and JAMIA). His papers have received around 15,000 citations and his H-index is 60. He has received numerous awards including NSF CAREER award, IEEE health informatics leadership award, as well as the awards from industries including Google faculty research award, Amazon machine learning for research award (twice) and Sanofi iDEA award. He has been the PI on numerous grants from federal agencies NSF, NIH, ONR and private foundations such as MJFF. Dr. Wang is a fellow of the American Medical Informatics Association (AMIA).

Marinka Zitnik, PhD, is an Assistant Professor at Harvard University with appointments in the Department of Biomedical Informatics, Broad Institute of MIT and Harvard, and Harvard Data Science. Zitnik investigates machine learning, focusing on graph representation learning, data fusion and knowledge graphs, few-shot learning, and their applications to network biomedicine and development of safe and effective therapeutics. Zitnik has published extensively in top AI/ML venues (e.g., NeurIPS, ICLR, ICML) and leading interdisciplinary journals (e.g., Nature Methods, Nature Communications, PNAS). She has organized numerous workshops and tutorials in the nexus of deep learning, drug discovery, and medical AI at leading conferences (NeurIPS, ICLR, ICML, ISMB, AAAI, WWW). Her research won Bayer Early Excellence in Science Award and numerous best paper and research awards from the International Society for Computational Biology. She was named a Rising Star in Electrical Engineering and Computer Science (EECS) by MIT and also a Next Generation in Biomedicine by the Broad Institute, being the only young scientist who received such recognition in both EECS and Biomedicine.

Shuiwang Ji, PhD, is currently an Associate Professor in the Department of Computer Science & Engineering, Texas A&M University, leading the Data Integration, Visualization, and Exploration (DIVE) Laboratory. He received the Ph.D. degree in Computer Science from Arizona State University in 2010. His research interests include machine learning, deep learning, data mining, and computational biology. Ji received the National Science Foundation CAREER Award in 2014. Currently, he serves as an Associate Editor for IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), ACM Transactions on Knowledge Discovery from Data (TKDD), and ACM Computing Surveys (CSUR). He regularly serves as an Area Chair or equivalent roles for AAAI Conference on Artificial Intelligence (AAAI), International Conference on Learning Representations (ICLR), International Conference on Machine Learning (ICML), International Joint Conference on Artificial Intelligence (IJCAI), ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), and Annual Conference on Neural Information Processing Systems (NeurIPS). Ji is a Distinguished Member of ACM and a Senior Member of IEEE.

Panel Organizer Statement: We affirm that all panel participants have agreed to participate and have contributed to the preparation of this document.
Stewardship Considerations in the Development and Implementation of Shareable SMART on FHIR Applications: Case Studies on Multiple Chronic Condition Care Planning and Chronic Pain Management

Laura Haak Marcial, PhD¹, Saira Haque, PhD, MHSA¹, Kensaku Kawamoto, MD, PhD, MHS², David A. Dorr, MD, MS³, Roland Gamache, PhD, MBA⁴; ¹RTI International, Research Triangle Park, NC; ²Department of Biomedical Informatics, University of Utah, Salt Lake City, UT; ³Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University, Portland, OR; ⁴Agency for Healthcare Research and Quality, Rockville, MD

Abstract

Healthcare management and treatment decisions are complex, with uncertain tradeoffs for patients and evolving evidence and guidelines for providers. For patients with multiple chronic conditions and those suffering from chronic pain, the challenges are more complex. This panel will discuss the development and implementation of three SMART on FHIR applications and will detail the development and implementation trade-offs and stewardship considerations needed to move from the sandbox environment to the test and production environments in these real-world settings.

The objectives of the presentation are to understand the complex challenges health systems face in integrating shareable SMART on FHIR applications, to evaluate the potential pitfalls and barriers to standards-based approaches for developing and implementing CDS using FHIR resources, and to assess the advantages for interoperability and knowledge sharing using the FAIR principles (findable, accessible, interoperable and reusable) and the CDS five rights based on the progress in the areas of CDS and FHIR based applications.

Background

Shareable clinical decision support (CDS) artifacts and other solutions that rely on access to fast healthcare interoperability resources (FHIR), such as those represented in the United States Core Data for Interoperability (USCDI), can be limited by the level of support and access to these FHIR resources. To describe these issues three real-world use-cases will be presented that highlight the implications of developing and implementing shareable Substitutable Medical Apps, Reusable Technologies (SMART) on FHIR solutions.

More than 25% of Americans have multiple chronic conditions (MCC), accounting for more than 65% of U.S. health care spending.¹ These individuals have complex health needs handled by diverse providers, across multiple settings of care. As a result, their care is often fragmented, poorly coordinated and inefficient. Therefore, easy data collection & retrieval (through eCare plans) is particularly important for people with MCC. Similarly, chronic (non-cancer) pain affects as many as one in three American adults² and one in ten American adults indicates this issue significantly disrupts their “work, social, and/or self-care activities.”³ These issues are linked to the use of opioids and the related co-morbidities and mortality of opioid use.⁴ The COVID-19 pandemic has further complicated MCC management and chronic pain and opioid management.⁵

The emergence of shareable interoperable solutions that leverage Health Level 7 International (HL7) standards like SMART on FHIR makes it possible to provide integrated solutions for improved care management. These shareable interoperable solutions can be extended with CDS by leveraging tools such as Clinical Practice Guidelines (CPG) on FHIR, CDS Hooks and clinical quality language (CQL). The Agency for Healthcare Research and Quality (AHRQ) has supported work on three projects:

1. The first project utilizes shareable, interoperable, and publicly available resources to address care plan management for patients with MCC through the development and pilot implementation of provider- and patient-facing tools for eCare plan management.
2. The second project leverages shareable, interoperable, and publicly available CDS tools, based on the CDC opioid guidance, to address chronic pain management with the development and pilot implementation of provider- and patient-facing tools to support shared decision making around chronic pain management.
3. The third project has developed an application for the translation of hypertension guidelines and protocols into tailored, interoperable, sharable CDS. The application supports providers and patients in making patient-centered
decisions about clinical actions based on how they interpret variation in conflicting guidelines for treatment in practice.

While the development of SMART on FHIR solutions is still emerging, this area shows a great deal of promise for enhancing focus and attention on disease management. Shareable CDS solutions are a novel way to build and propagate computable evidence-based practices in a variety of settings and EHR vendor systems. In practice, creating these interoperable tools has been limited to the sandbox environment where conditions are constrained. Moving these shareable solutions into real world environments has been limited to date. This panel discusses the real-world implications of developing and implementing shareable SMART on FHIR solutions to address multiple chronic conditions including chronic kidney disease, chronic pain, and hypertension and the challenges, opportunities, and ongoing management considerations which will need to be addressed to continue to expand into other contexts, vendor systems, and disease domains.

Learning Objectives

- Describe some of the complex challenges health systems face in integrating shareable SMART on FHIR applications with a focus on stewardship (ongoing management) and issues spanning multiple projects and implementing institutions.
- Provide key facilitators and barriers regarding stewardship of shareable standards-based approaches for developing and implementing CDS using FHIR resources.
- Assess the advantages, across projects and health systems, for interoperability and knowledge sharing using the FAIR principles and the CDS five rights based on the progress in the areas of CDS and FHIR based applications.

Speaker Contributions

- Dr. Gamache is a Staff Fellow at AHRQ. He will serve as the panel moderator and introduce the panel participants. To set the objectives of the panel, he will describe the major objectives of the work that AHRQ supports and promotes for shareable and interoperable CDS based on the CDS five rights and the FAIR principles.
- Dr. Kawamoto is Associate Chief Medical Information Officer of University of Utah Health. He will provide an overview of relevant standards including FHIR, SMART, CDS Hooks, and Clinical Quality Language (CQL). He will also provide examples of the use of these standards in a number of operationally deployed clinical applications.
- Dr. Marcial is a Health Informaticist at RTI International. She will highlight some of the implementation challenges associated with implementing the Clinical Decision Support for Chronic Pain Management (CDS4CPM) solution with specific reference to the key stewardship considerations identified in working with the health systems implementing shareable solutions.
- Dr. Haque is the Senior Director, Clinical Informatics Medical Outcomes Specialist at Pfizer. She will highlight experiences with implementation, use and evaluation of SMART on FHIR apps across organizations. She will focus on discussion of a patient-facing and provider-facing FHIR app to support e care plans for patients with MCC within and across organizations.
- Dr. Dorr is the Chief Research Information Officer at Oregon Health & Science University. He will discuss the process of implementation of these applications, the need for local validation and usability testing, and the evaluation process for effectiveness of these applications. He will focus on applications for high blood pressure, goal setting, and self-management support.

Expected Discussion and Discussion Questions

Panelists will stimulate discussions on issues of stewardship associated with shareable interoperable SMART on FHIR applications. Discussion questions include:

1. What kinds of resources will require special attention for long term (beyond the pilot period) of implemented shareable SMART on FHIR applications?
2. What are evaluation considerations for SMART on FHIR apps?
Attestation
The panel moderator has assurances from all participants that they will be available to participate at AMIA 2021.

References
Lessons learned the hard way from COVID-19: what gaps in laboratory informatics do we need to close before the next pandemic?

Patrick C. Mathias, MD, PhD\(^1\), Noah G. Hoffman, MD, PhD\(^1\), Brian R. Jackson, MD, MS\(^2\), S. Wesley Long, MD, PhD\(^3\), Amrom Obstfeld, MD, PhD\(^4\)

\(^1\)University of Washington, Seattle, WA, \(^2\)University of Utah, Salt Lake City, UT, \(^3\)Houston Methodist, Houston, TX, \(^4\)Children’s Hospital of Philadelphia, Philadelphia, PA

Abstract

Amidst significant barriers to scaling SARS-CoV-2 diagnostic testing during the pandemic, laboratories across the US developed a variety of solutions to expand access to testing and operate efficiently. In this panel, informaticists working in the domain of laboratory medicine will share their experiences scaling laboratories rapidly to address the needs of their communities and regions during the pandemic. Recurring themes in healthcare informatics will be discussed in the context of the pandemic: building vs. buying solutions, the benefits and drawbacks of HL7 as the standard mechanism to communicate orders and results, and the requirements for communicating actionable genomic information for infectious diseases. By exploring these themes, attendees will be able to consider developing and yet-to-be-developed informatics solutions to strengthen our healthcare and public health infrastructure.

Introduction

Throughout the first year of the COVID-19 pandemic few challenges were as prominent as scaling up and providing access to SARS-CoV-2 laboratory testing across the US. Clinical and public health laboratories across varying settings had to respond to an unprecedented need for testing amidst supply chain constraints, limited federal resources, and regulatory barriers. In the absence of clear solutions to manage the flow of orders, results, and data, laboratories performing SARS-CoV-2 testing were required to develop a variety of solutions to challenges in their setting and support the necessary response\(^1,2,3\).

The pandemic response re-introduced the common consideration of trade-offs in adopting the approach of building solutions vs. working with vendors to upgrade/purchase solutions and prompted organizations to ask: when is purpose-built custom software the best solution compared to an existing electronic health record system in meeting the needs for high throughput testing? Each of the organizations in this panel pursued different approaches in addressing the need to scale testing quickly and safely. Collecting and exchanging the information required for the public health response also emerged as an ongoing challenge. The traditional route of developing HL7 orders and results interfaces between laboratories and other entities often was not an agile approach at scale. These challenges clearly indicate the need for a more centralized approach to coordinating information flows, such as establishing a universal patient identifier, but did also allow collaborative problem-solving between laboratories and partners to evolve. In addition to navigating the challenges of providing accurate and rapid diagnostic services, disease surveillance focused on viral genomic sequencing emerged as an additional component of the broader public health response. The data generated by these activities largely exists and is exchanged outside of the standard data exchange mechanisms between laboratories and public health authorities, which has introduced an emerging need for informaticists to address.

In this panel, informaticists representing clinical laboratories across a variety of sizes and settings will discuss the challenges they faced in supporting the testing response and consider the current state of health information systems and the tradeoffs encountered in the context of an unprecedented pandemic. Each panelist will provide a short summary of the problems their organization needed to solve and discuss the solutions that were developed as well as the work that still needs to be done to improve the health of their population and of the public. By revealing different needs and challenges faced by a variety of organizations, the panel will consider the most prominent needs at the interface of laboratory services and the US public health infrastructure.
Panel Moderator:

**Patrick C. Mathias, MD, PhD** is an Assistant Professor in the Department of Laboratory Medicine and Pathology and an Adjunct Assistant Professor in the Department of Biomedical Informatics and Medical Education at the University of Washington School of Medicine. He serves as the Vice Chair of Clinical Operations and the Associate Medical Director of Informatics for the Department. His service work spans the configuration of the LIS and EHRs as well as the development of analytics and data science resources to support laboratory medicine and pathology quality improvement and research. As moderator, Dr. Mathias will outline the broad challenges of high-throughput test coordination and information exchange for laboratories and consider the remaining unsolved problems in which future development should be invested.

Panelist 1: Rapid implementation of high-volume community-based SARS-CoV-2 PCR testing in Washington State

**Noah G. Hoffman, MD, PhD** is an Associate Professor at the University of Washington and Head of the Informatics Division in the Department of Laboratory Medicine and Pathology. Dr. Hoffman is trained as a Clinical Pathologist, and his areas of specialization include medical informatics, bioinformatics, and software development for the clinical laboratory. Dr. Hoffman will describe the implementation and iterative improvement of the high volume SARS-CoV-2 PCR testing services developed in partnership between the University of Washington and local health authorities throughout Seattle, King County, and Washington State. Solutions to coordinating testing leveraged a combination of existing commercial systems, rapidly-developed custom applications and databases hosted both on premises and in the cloud, and third party practice management software for scheduling and data collection. The discussion will include a consideration of the organizational features and capabilities that contributed to the success of this collaboration, resulting in over one million test results delivered to date.

Panelist 2: Making an elephant dance: how the pandemic required a large national laboratory to adopt a startup mentality, to rapidly spin up new operational processes and associated software

**Brian Jackson, MD, MS** is Associate Professor of Pathology, and Adjunct Associate Professor of Biomedical Informatics, at the University of Utah. He is the medical director for IT, Business Development and Support Services for the University’s national laboratory business (ARUP Laboratories). Dr. Jackson will describe how COVID-19 testing required ARUP Laboratories to rapidly adopt new processes that were not anticipated within existing IT systems. In one use case, ARUP was asked to provide testing services for professional sports leagues as well as the 2020 U.S. vice presidential debate. This required development of an app for capturing patient demographics at the testing site and associating the patient id with a barcoded specimen. In another use case, ARUP received barcoded samples associated with orders captured by a third party system developed by a major software vendor unfamiliar with healthcare IT standards. This created a number of operational and IT challenges in order to reconcile this data in near-real-time for rapid testing.

Panelist 3: A wake up call for laboratory data interoperability

**Amrom Obstfeld, MD, PhD** is an Assistant Professor in the Department of Pathology and Laboratory Medicine at Children’s Hospital of Philadelphia and the Medical Director of Pathology Informatics for the Department. He serves as the primary liaison for informatic related issues spanning laboratory and clinical domains in addition to leading informatic and analytic projects within the Department. In this session Dr. Obstfeld will discuss the important role and significant challenges of interfacing SARS-CoV-2 test results during the coronavirus pandemic. Because of the critical nature of SARS-CoV-2 test results, the pandemic put a focus on the importance of health information exchange and data sharing between labs, hospitals and public health agencies. Nevertheless the urgent and rapidly evolving nature of the crisis often exacerbated long standing semantic and syntactic interoperability challenges which were superimposed upon a significantly deregulated laboratory environment. After systematically considering these barriers potential mitigation strategies will be discussed.

Panelist 4: Preparing for the new age of next generation sequencing based global pathogen epidemiology: How a hospital system in Houston became a leader in COVID-19 sequencing surveillance
S. Wesley Long, MD, PhD is an Associate Professor of Pathology & Genomic Medicine at Houston Methodist Hospital. He is the medical director of Diagnostic Microbiology and Pathology Informatics. Dr. Long will discuss how he prepared to leverage his research laboratory to sequence all of their hospital system’s COVID-19 isolates starting in January of 2020. As a part of this, he will discuss how they identified the ARTIC protocol and GISAID global network for COVID-19 epidemiology, how they leveraged and adapted existing pipelines for bioinformatics as well as protocols for sequencing. In particular, he will discuss the central tension between the need for local sequencing intelligence as well as the demands of public health authorities and global research groups. He will illustrate the importance of rapid iteration, nimble workflows, and overcoming supply chain issues to meet institutional goals as well as provide useful information to the global community to fight a growing pandemic.

Discussion Questions

1. What were the most significant obstacles laboratories, healthcare organizations, health jurisdictions, and communities faced in increasing access to testing?
2. What capabilities were required to rapidly develop solutions to scale testing to the necessary volumes and settings?
3. How should traditional healthcare informatics resources be reconfigured to respond rapidly to quickly changing needs?
4. What are the strengths and limitations of existing data exchange mechanisms between laboratories, healthcare organizations, and public health authorities?
5. How does data from new laboratory methods such as viral genomic sequencing fit into the broader landscape of health information exchange?

Learning Objectives

The panel will examine the current state and the future needs of providing laboratory services to healthcare organizations and the public to support the pandemic response as well as broader public health activities.

• Describe the standard architecture used to transmit laboratory data between clinical information systems.
• List the stakeholders involved and information sources required to communicate actionable information to public health jurisdictions.
• Compare the structure of emerging laboratory data types such as sequencing data with existing and developing standards for information exchange.

Target Audience

This panel will be of interest to informaticists working in the domains of information exchange and interoperability as well as those implementing EHRs to support population health scale efforts.

Panel Organizer Statement

All panelists have agreed to participate in this panel.

References

Informatics-izing the National Institutes of Health

PANELISTS

- Michael F. Chiang, MD, MA
  Director, National Eye Institute
  National Institutes of Health
- Patricia Flatley Brennan, RN, PhD
  Director, National Library of Medicine
  National Institutes of Health
- Joshua Denny, MD, MS
  Chief Executive Officer, All of Us
  National Institutes of Health
- Zhiyong Lu, PhD
  Deputy Director for Literature Search, National Center for Biotechnology Information (NCBI)
  Senior Investigator, National Library of Medicine
  National Institutes of Health

Moderator

- Clement McDonald, MD
  Chief, Health Data Standards Officer
  National Library of Medicine
  National Institutes of Health

ABSTRACT

The National Institutes of Health recognizes the importance of biomedical informatics innovations in accelerating data-driven discovery in biomedical science, clinical care and patient self-management. Of key importance is accessing information in the electronic health record. In the past three years, the NIH has expressed its public commitment to encourage researchers to adopt informatics innovations such as common data elements (CDEs), the Observational Medical Outcomes Partnership (OMOP) common data model, and standards such as the Fast Healthcare Interoperability Resource (FHIR) and the United States Core Data for Interoperability (USCDI). Policy statements are most effective when they are accompanied by operational procedures and vocal leadership direction. The purpose of this panel is to demonstrate the strength of informatics know-how among the leadership of the NIH, identify existing commitments by the NIH to ensure accelerating the research value of clinical health records, and to glean insights of strategies the NIH should consider as they develop the informatics infrastructure for the future of biomedical discovery.

DESCRIPTION OF THE PANEL

For the first time in its 133-year history, the National Institutes of Health (NIH) has major biomedical informatics leadership. The NIH recognizes that the future of biomedical discovery rests in part on being able to leverage the knowledge embodied in clinical information records. For example:
More so than ever before, understanding the natural history of a pandemic and ensuring best practices in being able to treat it relies on gleaning important insights from clinical records. More importantly, the ability to quickly and efficiently access clinical information provides an opportunity to titrate clinical trials in response to in-the-moment understandings of the course of illness and response to care for patients.

Better understanding of the long-term course of the pandemic and its clinical sequelaes rests on being able to track patients across time. Clarifying the impact of novel vaccines or clinical therapeutics would be enhanced by being able to integrate participant information across any and every study in which the participant is represented.

There is increasing recognition about research that emphasizes the re-usability and machine-actionability of large-scale data for applications such as machine learning and artificial intelligence. At the same time, there is appreciation for the importance of interdisciplinary approaches to collaboration across diverse social, cultural, economic, academic, and industrial backgrounds and communities.

The NIH is deeply engaged in exploring the value of contemporary and emerging informatics innovations including cloud based reusable platforms/datasets, common data models, and the effective development and use of artificial intelligence and machine learning approaches to biomedical research and clinical practice.

Each of these requires effective deployment of informatics innovations into the research process.

AMIA’s Clinical Research Informatics community has done much to enhance the integration of biomedical informatics concepts into the research process. Much progress has been made in the efforts to structure information for clinical and translational research and common data elements. Incremental progress must be praised, but to engage the operations of the world’s largest biomedical enterprise requires multiple points of engagement. Specifically, expanding the critical mass of leadership with expertise in biomedical informatics is essential for instituting enterprise-wide change.

The National Library of Medicine (NLM) has long served as the voice of biomedical informatics at the NIH. Innovations such as the Unified Medical Language System and the development of innovative clinical information systems to extract clinically specific phenotypes served the biomedical enterprise. The NIH has adopted many of the NLM-advocated strategies for data alignment and information reuse, including common data elements. The next step in the process is to engage the NIH in an enterprise commitment to the use of clinical data standards (USCDI, FHIR). The final step is engagement in the operational level of the NIH enterprise, and that requires instantiation at the highest levels of NIH leadership – the ability to translate the operational goals of the various institutes and centers (ICs). This requires people knowledgeable in biomedical informatics to sit at the table with the directors across all of the NIH.

Perhaps as evidence of the importance of biomedical informatics in the research enterprise, NIH now has directors of three institutes and/or major operations who hold standing in the biomedical informatics community. The number of American College of Medical Informatics fellows among the NIH staff is expanding. Not only does this allow a ‘community of conversation’ but embedding informatics expertise across the NIH has accelerated the acceptance of, and the valuing of, biomedical informatics in the biomedical research enterprise.

In this panel, three NIH Directors with background in biomedical informatics, as well as an experienced NLM intramural investigator, will discuss perspectives on these topics and how they are evolving at NIH.
DISCUSSION: CHALLENGES AWAIT US

The panel will end with a discussion period involving panelists and audience members. This will focus on collecting insights from participants about approaches for the NIH to consider in developing informatics infrastructure for the future of biomedical discovery. Special characteristics of the NIH, particularly the diversity of research cultures across 27 institutes and centers, do not lend themselves to highly orchestrated solutions. Design thinking, which involves engaging both program staff and the extramural community, will more likely lead to solutions that allow at once for local specification while promoting trans-NIH integration. Additionally, any solutions proposed by the NIH will need to align with business and clinical operations in over 3000 health care delivery systems; human factors approaches that support specification of operational practices and leverage rather than disrupt local cultures will only arise through effective analysis and modeling. This panel is designed as much to share the NIH perspective as it is to glean wisdom and guidance from the AMIA community.

In summary, the NIH investigators are ready to engage with clinical data for the purposes of research, but many challenges remain. The first challenge rests in arriving at a common nomenclature to address the broad range of clinical data. The second challenge must address the problem of interoperability for research – and how do we ensure that data generated in the course of care become useful for clinical research? Additional concerns arise in how the original consent for data collection and whether future research reuse considerations may be adequately addressed at the point when consent is obtained.

REFERENCES

Biomedical Informatics Co-design: Concepts, Applications, and Opportunities

Mollie McKillop, PhD, MPH\(^1\), Anne Moen, RN, PhD, FACMI, FIAHSI\(^2\), Rosemary Kennedy PhD, RN, MBA, FAAN\(^3\), Andrew Berry, PhD\(^4\), Rupa Valdez, PhD\(^5\)

\(^1\)IBM Watson Health, Cambridge, MA, USA; \(^2\)University of Oslo, Oslo, Norway; \(^3\)Connect America, Bala Cynwyd, PA, USA; \(^4\)University of Washington, Seattle, WA, USA; \(^5\)University of Virginia, Charlottesville, VA, USA

Abstract

Individuals whose activities and experiences will be most impacted by a new technology should have input into the design process. These individuals offer invaluable expertise with respect to perspectives and preferences for the workflow or chain of activities the technology is intended to impact. How well a technology fits into a specific workflow influences its adoption and use. Collaborative design is increasingly important to capture and better reflect individualized activities and experiences. At the same time, there are pragmatic challenges such as participant trust, recruitment, and engagement in the co-design process. Moreover, the COVID-19 pandemic has pushed researchers to find novel and creative ways of co-designing with end-users. In this panel, we will discuss co-design core concepts and research trends in biomedical informatics through practical examples from the research experiences of a diverse set of panelists.

Introduction

Designing health information technologies with patients, providers, and citizens (co-design), who are the intended end-users of these solutions, generally makes them more acceptable and usable. The value of co-design is that it can highlight the sociotechnical and cultural aspects of clinical care and patient self-management that may ultimately change behaviors and improve outcomes\(^1\). Approaches to identifying these aspects include participatory research design (community collaboration and engagement across all research stages), participatory design (focused and purpose driven collaboration with end users), and user-centered research (participation around specific design features). Given these differing approaches, there remains an opportunity for better clarity on the different ways of engaging with patients, caregivers, and the community for co-design. Challenges in this area include, for example, the effort required to build partnerships with relevant communities, especially if the medical establishment has historically marginalized them. In addition, many informatics tools currently in practice are limited in realizing the value of co-design by their focus on technology and deterministic evaluation outputs rather than the social or psychological aspects of users they intend to impact. In this panel, we explore concepts, ideas, and approaches to co-design across a wide variety of populations to identify common methodological themes and opportunities. For example, co-design may be more useful when it is seen as part of an iterative development process for generating new knowledge about a phenomenon rather than a method to achieve a specific outcome metric.

We will focus on three population use cases in co-design: marginalized and hard-to-reach populations, the elderly, and those with disabilities. The patient or caregiver perspective will be emphasized. We plan to talk about the role of co-design in technology-based interventions that span the trajectory from independence to emergency care to longer term care needs. Approaches and challenges in co-design methods across these use cases will be examined. For example, we will discuss the development of in-home technology and care delivery models to support aging safely in place. Ways of eliciting cancer care patient needs for integrative oncology services will also be covered.

At the end of this panel session, attendees can expect to better understand: (1) what co-design is and how it may best be viewed and applied; (2) aiding selection of co-design methods, (3) learning how to engage diverse groups in co-design, and (4) leveraging multiple modes of user engagement.

All panelists have agreed to take part on the panel.

This panel is endorsed by the AMIA Consumer and Pervasive Health Working Group Leadership.
Multimodal Co-creation for Personal Health Information Management

Active engagement in personal health activities requires access to actionable, understandable, relevant, reliable, and up-to-date, evidence-based information that meets the user’s specific health context and literacy level. However, lack of well-designed tools to enable such access is a significant problem for efficient personal health information management support across activities for a) episodic self-care, b) during complex treatment and frequent interactions with the health system, or c) continuous self-management. We will contribute experiences from co-creation workshops with video material, interviews and experiences to represent user voice to design accessible and universally designed digital tools where the users can 1) collect relevant health information, 2) complement or annotate according to personal, situational need, 3) collaborate with trusted partners and the health care team, and 4) choose to share but retain control of “what to share, with whom, and for how long”. A shared concern is to strike the best balance of digital health literacy capacity, availability of information for collaboration, integrity, confidentiality and personal preferences for trust, and legal requirements for data protection.

Co-design Methods, Opportunities, and Challenges in Mobile App Development

The explosion of digital health technologies has challenged traditional institution-centric care. Virtual care has rapidly evolved, creating reciprocal relationships between patients and providers where patients are involved in the production process. For example, in telehealth, patients use videoconferencing, document symptoms, and capture data using digitally enabled medical devices. For this reason, the people using the solution may be best positioned to help create it through the use of co-design principles. There are opportunities to use co-design methods when developing digital health solutions and care delivery models to share responsibility and achieve value for all stakeholders. Methods and findings from a virtual co-design study will be presented. The co-design study focused on identifying patient needs for integrative oncology services as a foundation for mobile application development. Opportunities and challenges associated with the use of co-design methods will be explored using case examples from the study.

Choosing and Chaining Co-design Methods: Concepts and Concrete Examples

What should researchers and practitioners consider when choosing co-design methods for a particular problem, context, and set of stakeholders? This presentation offers guidance on bringing together multiple co-design methods to increase the depth and rigor with which design ideas are generated, elaborated, and evaluated. This guidance will be grounded in a case study of a series of future workshops, storyboarding workshops, and focus groups that characterized the design space for improving patient-provider communication about patients’ personal values and health priorities. Analyzing the case studies through the lens of design theory on hybrid spaces4 and inventive methods5, as well as standards for high quality qualitative health research (credibility, transferability, dependability, and confirmability6), this presentation offers theoretical concepts and concrete examples to guide researchers and practitioners in matching co-design methods to the design context.

Participatory Research with Underserved Populations

Challenges and lessons-learned related to scoping a joint research focus, building long-standing relationships, and engaging in participatory interactions using virtual platforms, particularly as related to underserved populations will be discussed. Negotiating publications and traditional academic outputs in the context of such partnerships will also be covered. Attendees should be able to describe the challenges associated with participatory design-based approaches to informatics research in home and community-based settings and associated strategies to overcoming them.

Moderator and Panelist Biographies

Mollie Mc Killop, PhD, MPH (panel moderator) is a biomedical informatician at the Center for AI, Research and Evaluation (CARE). In this role, Mollie provides technical, informatics, and evaluative leadership in the design, development, and execution of research evaluation studies across IBM Watson Health, focusing on consumer-facing offerings in the payer and government market segments. Mollie is also the publications and publicity co-chair for the AMIA Consumer and Pervasive Health Working Group.

Anne Moen, RN, PhD, FACMI, FIAHSI (panelist) is professor at the Faculty of Medicine, University of Oslo, Oslo,
Rosemary Kennedy PhD, RN, MBA, FAAN (panelist) is Chief Health Informatics Officer at Connect America. She is a recognized informatics domain expert holding many leadership roles through her work with the AMIA and Technology Informatics Guiding Educational Reform Board. Dr. Kennedy is widely presented and published in the field of nursing informatics, clinical documentation, and terminology standards. She is a fellow in the American Academy of Nursing and a recipient of the HIMSS Nursing Informatics Award and Modern Healthcare Top 25 Women in Healthcare Award. She served as Vice President of Health IT for the National Quality Forum in Washington DC and Chief Nursing Informatics Officer for Siemens Healthcare Solutions.

Andrew Berry, PhD (panelist) is a design researcher focused on improving health equity for people with complex health and social needs. He is a National Library of Medicine Postdoctoral Fellow in Biomedical Informatics with a Ph.D. in Human Centered Design & Engineering and applied industry experience in technical services at Epic. He uses a mixed-methods approach, with a focus on qualitative, human-centered, and participatory methods for research and design.

Rupa Valdez, PhD (panelist) is an Associate Professor of Biomedical Informatics and Engineering Systems and Environment at the University of Virginia. Her research draws heavily on participatory design and community-engaged approaches to understand and design systems to support patients and their social networks (e.g. family, friends, neighbors) with managing health at home. Much of this work focuses on underserved populations, including racial and ethnic minorities, people with disabilities, and people living in under-resourced settings.

Discussion Questions
1. What does co-design mean to you?
2. How have you built trust with the communities you work with?
3. What excites you the most about community collaboration, co-design, or participatory research?
4. What do you think the future of the field of co-design in informatics looks like?
5. How do you choose which co-design methods are appropriate for which applications or populations?
6. For those who are new to the field, what would you recommend for both career development and building expertise?
7. How do you think we can move the field forward so that more informatics tools include a co-design development approach that is inclusive and useful for all?
8. What are the biggest challenges you face in co-design research and how have you worked to overcome these challenges? What are the biggest opportunities in the field?

References
6. Devers, K. J. (1999). How will we know “good” qualitative research when we see it? Beginning the dialogue in health services research. Health Services Research, 34(5 Pt 2), 1153–1188.
Balancing Access to Federal Data Sets with Enhancing Privacy and Security

Lisa B. Mirel, M.S.1, Stephanie Garcia, M.P.H.2, Stefan Jaeger, Ph.D.3, Ziv Yaniv, Ph.D.4, Patricia Keenan, Ph.D.5, Susan Lumsden, M.S.6

1 Centers for Disease Control and Prevention, U.S. Department of Health and Human Services, Atlanta, GA; 2Office of the National Coordinator for Health Information Technology, U.S. Department of Health and Human Services, Washington, DC; 3National Library of Medicine, U.S. Department of Health and Human Services, Bethesda, MD; 4National Institute of Allergy and Infectious Diseases, U.S. Department of Health and Human Services, Bethesda, MD; 5Agency for Healthcare Research and Quality, U.S. Department of Health and Human Services, Rockville, MD; 6Assistant Secretary for Planning and Evaluation, U.S. Department of Health and Human Services, Washington, DC;

Abstract
The Office of the Secretary Patient-Centered Outcomes Research Trust Fund (OS-PCORTF) portfolio is charged with building data capacity for the conduct of patient-centered outcomes research (PCOR). Federal data sets may contain sensitive information, such as patient-level identifiers. Therefore, access to federal data sets for research is often limited due to privacy protection concerns. This panel features four OS-PCORTF projects in the Department of Health and Human Services’ (HHS) that balance access to data with privacy and security concerns, including privacy preserving record linkages between federal data sets, synthetic data to eliminate the risk of re-identifying anonymized data, developing training data sets to improve research capacity for machine learning, and developing publicly available databases with social determinants of health (SDOH) data at the small-area level. Panelists will each provide an overview of their project, describe the contributions to ensure high-quality health care-related data, and discuss informatics-related opportunities to protect privacy.

Learning Objectives
1. Learn methods to protect privacy when one needs access to data to conduct PCOR.
2. Identify contributions by the federal government on the use of informatics to support PCOR.

Panel Description
This panel will highlight four OS-PCORTF collaborative projects focused on an overarching goal of developing and assessing promising methods to protect privacy while making federal data sets more accessible. Specifically, the projects focus on (1) using privacy preserving record linkages between federal data sets; (2) using synthetic data to eliminate the risk of re-identifying anonymized data; (3) developing training data sets to improve the research capacity for machine learning; and (4) developing a consolidated set of databases with social determinants of health (SDOH) factors at the small-area level as opposed to the individual-level.

Lisa B. Mirel, M.S., from the Centers for Disease Control and Prevention (CDC), National Center for Health Statistics (NCHS), will speak about her work on the project, “Data Linkage: Evaluating Preserving Privacy Methodology.” The project aims to conduct a methodological assessment of privacy preserving record linkage techniques (PPRL). The National Hospital Care Survey data will be linked to the National Death Index using PPRL techniques and then compared to the linkage results using a standard linkage algorithm that rely on the exchange of unencrypted identification data. The project will assess how the use of PPRL may affect the quality of the linked data (precision and recall) and further affect the inference (estimates) in secondary analysis of those data. This assessment will inform the health data ecosystem on the evaluation of PPRL methodology and may increase the linkage activities at NCHS.

Stephanie Garcia, M.P.H., from the Office of the National Coordinator for Health Information Technology (ONC), will present the project, “A Synthetic Health Data Generation Engine to Accelerate Patient-Centered Outcomes Research.” High-quality health care-related data are often difficult to access because of cost, patient privacy, or
other legal and intellectual property restrictions. As conceptualized in another OS-PCORTF project, “Privacy and Security Blueprint, Legal Analysis and Ethics Framework for Data Use, and Use of Technology for Privacy,” researchers and developers often depend on anonymized data to test theories, data models, algorithms, or prototype innovations to protect privacy. However, the risk of re-identification of anonymized data is high, especially with rare conditions. Further, due to a variety of interoperability issues, it is often difficult to bring data together from different resources for the purpose of robustly testing analysis models, algorithms, or assisting in the development of software applications. Use of synthetic data eliminates the risk of re-identifying anonymized data—which can occur when using de-identified datasets—and bypasses the interoperability challenges that can stem from combining disparate data sources for PCOR. The project will enhance the capabilities of Synthea™, an open-source software that creates large amounts of clinically realistic, synthetic patient health records. The project will increase the number of synthetic patient health records by developing data generation modules for opioid, pediatric, and complex care use cases.

Stefan Jaeger, Ph.D., from the National Library of Medicine (NLM) will present work he conducted with Ziv Yaniv, Ph.D., from the National Institute of Allergy and Infectious Diseases, on an interagency project between NLM and ONC titled “Training Data for Machine Learning to Enhance Patient-Centered Outcomes Research (PCOR) Data Infrastructure.” Artificial Intelligence (AI) and associated innovative technologies like machine learning consume large amounts of data in varied, complex formats to identify effective treatments, potentially accelerating clinical innovation by speeding up the research lifecycle and the application of evidence in clinical settings. Machine learning relies on training data sets and training algorithms to “learn” how to identify patterns with little human intervention. The project will improve research capacity for machine learning by curating high-quality training data sets. The training data sets will use HHS research data, including linked clinical data. The project will develop a use case for detecting drug resistance in tuberculosis patients, leveraging federal data assets from the National Institute of Allergy and Infectious Diseases (NIAD).

Patricia Keenan, Ph.D., from (AHRQ), will present AHRQ’s work on the project, “Enhancing Patient-Centered Outcomes Research (PCOR): Creating a National Small-Area Social Determinants of Health (SDOH) Data Platform.” For many decades, researchers have emphasized the importance of SDOH and have developed many different conceptual models to explain the inter-relationship between individual, family, and societal factors on the health of an individual. However, SDOH data may not always be easy for providers to collect at the patient level and patients may not always want to share individual-level SDOH data. This project makes SDOH data available for public use by developing a consolidated set of publicly available national standardized databases on valid and reliable SDOH factors at the small-area and other geographic levels. The project is building on existing federal databases maintained by agencies such as AHRQ, Health Resources and Services Administration (HRSA), CDC, Assistant Secretary for Planning and Evaluation (ASPE), and the National Institutes of Health (NIH). Data elements will span the SDOH landscape and include measures of social context, economic context, education, physical infrastructure, and healthcare context.

Following the presentation, Ms. Susan Lumsden will moderate a Q&A between the panelists and audience members.

**Timeliness of Topic**

To answer complex research questions aimed at improving value and whole-person care requires the ability to link data sets and aggregate data from disparate sources. However, data breaches erode trust among patients and participants to share their data and have huge financial implications. In 2019, 64% of Americans were concerned about how the government uses the data collected. In 2020, the average cost of a data breach in the healthcare industry was $7.13 million. The risk of data breaches emphasizes the importance of measures that protect privacy among clinical and federal data sets. There is a need to balance access with enhancing privacy and security so that researchers can access representative data sets to conduct robust PCOR. The projects featured on this panel contribute to the availability of federal data or leverage federal data to offer alternatives, which can ultimately improve PCOR and ultimately health outcomes. The panel will hopefully spur more strategies that enhance privacy and security and inform how research and health care entities can better balance data access with security.

**Discussion Questions**

1. How have interagency collaborations enhanced the work in privacy preservation?
2. What are real world applications of the methods from each project?
3. What are next steps for the projects and how will findings and resources be disseminated?
All participants have agreed to take part in the proposed panel.

References


Shawn N. Murphy, MD, PhD1,2, Hossein Estiri, PhD1,2, Arianna Dagliati, PhD3, Riccardo Bellazzi, PhD3,4, John H. Holmes, PhD5

1Lab of Computer Science, Massachusetts General Hospital, Boston, MA; 2Harvard Medical School, Boston, MA; 3Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Italy; 4LISRC Lab, ICS Maugeri Hospital, Pavia, Italy; 5University of Pennsylvania Perelman School of Medicine, Philadelphia, PA

Abstract

While the fight against the SARS-CoV-2 virus has ramped up with new vaccines and therapeutics, there is a growing urgency to study and address another side of the pandemic, the shapeshifting syndrome often known as the post-acute sequelae of COVID-19 (PASC). This panel is designed to share the experiences and innovative approaches of the Consortium for Clinical Characterization of COVID-19 by EHR (4CE) to study and model the PASC across an internationally diverse group of healthcare institutions. The panel will discuss the organization of 4CE, computational phenotyping to identify cases of post-COVID syndromes, the identification of temporal phenotypic pathways, and the application of sociobehavioral agent-based network models to explore those pathways for prediction and intervention in population contexts.

Introduction and Background

A collection of persistent physical (e.g., fatigue, dyspnea, chest pain, cough), psychological (e.g., anxiety, depression, post-traumatic stress disorder), and neurocognitive symptoms (e.g., impaired memory and concentration) can appear and last for weeks or months in patients after acute COVID-19.1–3 These Post-Acute Sequelae of SARS-CoV-2 infection could impede the ability of some coronavirus patients to work and function normally,4–7 costing the U.S. economy billions of dollars annually in medical bills and lost incomes for an unforeseeable future. While the exact number of people afflicted by post-Covid syndrome is unknown, it represents a significant public health burden because of the large magnitude of the COVID-19 spread globally. Further, the extent to which the post-acute sequelae of COVID-19 represents a new syndrome unique to COVID-19, or an overlap with the recovery from similar illnesses, is unknown. Furthermore, little is known about COVID-19 phenotypes and subphenotypes, especially in a temporal context and how these phenotypes emerge in sociobehavioral contexts.

In March 2020 an international effort called the Consortium for Clinical Characterization of COVID-19 by EHR (4CE) was launched. It used a federated model, where local experts in informatics, statistics and clinical medicine were engaged at more than 340 hospitals across 9 countries on 4 continents to run analyses locally at each site (sharing only aggregate results) and to iteratively improve the data quality in the local databases. The 4CE network was created to study EHR-derived data from multiple institutions and was based upon experience that it can only be successful if you trust the findings. This means being able to go back to each site and talk to the people who know the data best, to develop and refine the research questions and understand the local clinical guidelines, coding practices, data quality problems, and other factors that affect the data, and compare it to the EHR directly though Chart Review.

Panel Description
Dr. Shawn Murphy will lead a panel discussion on the Post-Acute Sequelae of SARS-CoV-2 Infection across the 4CE consortium, and in doing so will help the audience better understand the possibilities of leveraging EHR data to identify the phenotypes and its subtypes within the U.S. and internationally. Dr. Murphy is a Professor of Neurology and Biomedical Informatics at Harvard Medical School, Chief Research Information Officer at Mass General Brigham, and the Associate Director of the Lab of Computer Science at MGH.

The panel will include a review of studies across the 4CE consortium to understand temporal patterns of analysis that are able to tease out events sequences that allow observing the unique characteristics of a disease as it progresses. The panel will also discuss the future implications of international collaborations to address the post-Covid sequelae using large scale clinical data.

Given the global public health significance of the Post-Acute Sequelae of SARS-CoV-2 infection and the international outreach of the 4CE consortium, the proposed panel is well suited for the AMIA members and audience.

**Presenters**

**Post-Acute Sequelae - the Italian case (Bellazzi)**

Dr. Riccardo Bellazzi is Professor of Bioengineering and Medical Informatics, Director of the Department of Electrical, Computer and Biomedical Engineering of the University of Pavia, and director of the LISRC labs of the ICS Maugeri hospital in Italy. Starting from the experience coming from the 4CE initiative, in which we collected EHR data coming from more than 4,100 italian patients, Prof. Bellazzi will discuss the organizational aspects related to the data collection and curation of long-Covid patients highlighting differences and similarities between the Italian and US case, and he will also address some computational challenges of long-COVID phenotyping.

**Computational Phenotyping to Identify Post-Covid Syndrome in Outpatient Population (Estiri)**

Dr. Hossein Estiri is an assistant professor of medicine at MGH lab of computer science and Harvard Medical School. Many of the post-Covid symptoms could have multiple causes. Yet, the extent to which the post-acute sequelae of COVID-19 represents a new syndrome unique to COVID-19, or an overlap with the recovery from similar illnesses, is unknown. In his presentation, Dr. Estiri will describe application of computational phenotyping for identifying the clinical records that positively associate with a patient having tested positive for COVID-19 in the past. He will introduce approaches for effectively leveraging large scale clinical data to study and model the post-acute sequelae of COVID-19 (PASC). Finally, Dr. Estiri will present novel diagnosis and biomarkers of the PASC identified through the application of the computational phenotyping to COVID-mart clinical data from the Mass General Brigham in Boston.

**Temporal Trajectories and Careflows models to analyse Post-Covid clinical pathways (Dagliati)**

Dr. Arianna Dagliati is an assistant professor of Bioengineering and Medical Informatics at the Department of Electrical, Computer and Biomedical Engineering of the University of Pavia. Patients affected by post-Covid symptoms might undergo heterogeneous clinical courses both from the physiopathological and process-of-care points of view. Temporal phenotyping based on discrete events sequences allows observing the characteristics of a disease as it progresses, especially in relation to care processes. Other approaches
capture interactions between disease changes over time from continuous multivariate data and define trajectories as key features to characterize disease subtypes. Using data from two Italian hospitals in the Lombardy region, Dr. Dagliati will describe the application of temporal phenotyping methods to derive post-Covid phenotypic paths from multivariate trajectories and sequential patterns, and how the integration of these approaches might enable clinicians to identify early warning signs of progression in specific post-Covid patients.

**Modeling Post-COVID Chaos: Dynamic Non-linear Models of Post-Covid Trajectories for Hypothesis Exploration (Holmes)**

Dr. John H. Holmes is Professor Medical Informatics in Epidemiology in the Department of Biostatistics, Epidemiology, and Informatics and Associate Director of the Institute for Biomedical Informatics at the University of Pennsylvania Perelman School of Medicine. He specializes in novel machine learning approaches to mining biomedical data and in agent-based simulations of temporal phenomena. Using data from the University of Pennsylvania Health System, Dr. Holmes will present sociobehavioral agent-based modeling methods to explore various temporospatial scenarios affecting patient trajectories as approaches to predicting and intervening on post-COVID sequela.

**Discussion Questions**

1. What are the Post-Acute Sequelae of SARS-CoV-2 Infection?
2. What are the differences in COVID post-acute sequelae between the U.S and Italy?
3. How can computational phenotyping be applied to define the post-COVID phenotypes?
4. How can heterogenous longitudinal data be used to explain post-COVID syndrome and its progression?
5. How can computational phenotypes be integrated into clinical care for post-covid follow up?

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

**References**


Laurie L. Novak, PhD, MHSA, FAMIA¹, Carla Perissinotto, MD², George Demiris, PhD, FACMI³

¹Vanderbilt University Medical Center, Nashville, TN; ²University of California San Francisco, San Francisco, CA; ³University of Pennsylvania, Philadelphia, PA

Abstract

Social isolation and loneliness are underappreciated health risks associated with higher rates of mortality, depression, and cognitive decline, particularly in older adults, and are exacerbated by the COVID-19 pandemic. This panel discusses implications of the National Academies of Science, Engineering, and Medicine, report entitled “Social Isolation and Loneliness in Older Adults: Opportunities for the Health Care System.” The report outlines numerous roles for informatics in addressing this public health challenge, including infrastructure for coordination among community providers, and clinical and community interventions. The panel will review the recommendations and their relationship to social determinants of health, discuss how the pandemic has intensified and expanded the public health risks related to social isolation and loneliness, and how informatics may play a role both in addressing risks and creating resilient systems for future pandemic threats. The audience will be engaged in a discussion to identify strategies for effective implementation of the recommendations.

Panel Description

Social isolation (an objective lack of social contact with others) and loneliness (the subjective feeling of being isolated) are significant yet underappreciated public health risks. Approximately one-quarter (24 percent) of community-dwelling Americans aged 65 and older are considered to be socially isolated, and a significant proportion of adults in the United States report feeling lonely (35 percent of adults aged 45 and older and 43 percent of adults aged 60 and older).¹⁻³ Research has shown that social isolation significantly increases a person’s risk of mortality from all causes, a risk that may rival the risks of smoking, obesity, and lack of physical activity.⁴ Social isolation has been associated with a 29 percent increased all-cause risk for mortality and a 25 percent increased risk for cancer mortality.⁵,⁶ Loneliness has been associated with a 59 percent increased risk of functional decline and 45 percent increased risk of death.³

With support from the AARP Foundation, the National Academies of Sciences, Engineering, and Medicine Committee on the Health and Medical Dimensions of Social Isolation and Loneliness in Older Adults issued a report in February, 2020.⁷ The committee’s charge consisted of two parts: 1) to examine how social isolation and loneliness affect health and quality of life in adults aged 50 and older, particularly among low-income, underserved, and vulnerable populations, and 2) to identify opportunities for clinical settings of health care to help reduce the incidence and adverse impacts of social isolation and loneliness. As an outcome of the study, the committee outlined five goals, each with implications for informatics:

1. Develop a more robust evidence base for effective assessment, prevention, and intervention strategies for social isolation and loneliness;

2. Translate current research into health care practices in order to reduce the negative health impacts of social isolation and loneliness;

3. Improve awareness of the health and medical impacts of social isolation and loneliness across the health care workforce and among members of the public;

4. Strengthen ongoing education and training related to social isolation and loneliness in older adults for the health care workforce; and

5. Strengthen ties between the health care system and community-based networks and resources that address social isolation and loneliness in older adults
Learning Objectives

After attending this panel audience members will better be able to:

1. Describe the health impacts of social isolation and loneliness in older adults.
2. Describe and assess technology’s role in mitigating the health impacts of social isolation and loneliness using electronic health records and other technology-based interventions.
3. Discuss what COVID-19 has taught us about older adults and barriers to technology use and what safeguards can be put in place to create more resilient health systems.

The panel comprises three members of the NASEM committee. Panelists will review the report and the recommendations and describe how they fit within the larger context of social determinants of health. The panelists will then engage the audience in a discussion to identify strategies for effective implementation of the recommendations.

Dr. Novak will moderate the panel, introducing the topic and discussing how existing and emerging technology infrastructure can support the ecosystem of health care and community providers in implementing report’s recommendations. She will also describe the potential role of predictive analytics in identifying populations at risk, and the related social and practical implications.

Dr. Perissinotto will review the implications for technology used in clinical settings. She will specifically focus on how assessments can be integrated into electronic medical records, including choice of tools, the use of discrete data fields and potential barriers to implementation. She will review lessons from the COVID-19 pandemic related to older adults and technology use.

Dr. Demiris will describe committee findings related to technology interventions (including mHealth applications, conversational agents, other artificial intelligence and social connectivity tools) and current evidence as to the effectiveness of such systems. He will also discuss ethical implications associated with the design and implementation of informatics tools to address loneliness and social isolation in community settings and recommendations for system designers of future tools.

Relevance

As health care organizations continue to refine the implementation of electronic health record (EHR) tools to monitor and influence the social determinants of health, it is necessary to engage informaticists and clinicians in dialogue to ensure informed deployment and use of the tools. This NASEM report and its implications for a range of technologies and clinical practices is a catalyst for the informatics community to engage in such a dialogue. We anticipate this panel to be of interest to data scientists, clinicians, clinical informaticists, and health care operational and IT leaders.

Discussion Questions

1. What technology infrastructure innovations can accelerate effective coordination between health care providers and community service organizations?
2. What are the barriers to implementing effective assessments in various types of clinical settings?
3. What are the best uses of mHealth for reducing social isolation and loneliness?

Panel Participants

Laurie Lovett Novak, PhD, MHSA

Dr. Novak is Assistant Professor and Director of the Center of Excellence in Applied Artificial Intelligence in the Department of Biomedical Informatics (DBMI) at Vanderbilt University Medical Center. Her research aims to improve the beneficial alignment of technology and social life. Current work is focused on developing and implementing AI tools in a variety of clinical environments and creating new ways of understanding and supporting patient and caregiver management of chronic illness in everyday life. Dr. Novak is an expert in designing and conducting research on the clinical work activities of patients, caregivers, and clinical personnel. She serves as a social science resource for DBMI, advising and teaching others on topics such as qualitative methods of research,
formulation of research questions that relate to technology, work, and society, and facilitation of research and design partnerships between computational scientists and clinicians.

**Carla Perissinotto, MD**

Dr. Carla Perissinotto is the Associate Chief for Geriatrics Clinical Programs at UCSF and Medical Director for the UCSF Care at Home practice which provides geriatrics and palliative care to homebound adults. In her roles, Carla strategically oversees and manages both inpatient and outpatient clinical programs which include Care at Home, Geriatrics consultations, the Acute Care for the Elderly (ACE) unit at Moffitt-Long Hospital, Orthopedic Co-Management programs, the surgical Prehab clinic, and community partnerships. Carla is an Associate Professor in the Division of Geriatrics, Department of Medicine. Carla is Board Certified in Internal Medicine, Geriatrics and Palliative Medicine. Carla is dedicated to working in both community and academic settings that serve diverse patients. From 2008-2017, Carla served as a primary care and consulting Geriatrician at the Over60 Health Center in Berkeley. Carla has also gained national and international recognition for her research and advocacy on the effects of loneliness on the health of older adults and is currently focused on academic-community partnerships to understand implementation of loneliness interventions.

**George Demiris, PhD, FACMI**

George Demiris is a PIK (Penn Integrates Knowledge) University Professor in the School of Nursing with a joint appointment in the Department of Biostatistics, Epidemiology and Informatics in the Perelman School of Medicine at the University of Pennsylvania. His research focuses on the use of information technology to support older adults and their family caregivers and explore innovative solutions to promote independent aging and patient and family engagement. He is a co-founder of the Hospice Caregiver Research Network, an initiative led by researchers from various academic disciplines committed to designing and testing interventions to support family caregivers of patients at the end of life. Another area of his research includes the use of behavioral sensing, “smart home” and “Internet of Things” technologies to promote independence for community dwelling older adults and their families. Such emerging technologies introduce challenges and opportunities in terms of engaging older adults in decision making, making sense of vast amounts of data and promoting effective data visualizations as well as addressing ethical considerations.

**References**

Stakeholder-driven “Art of the Possible” Patient Journeys for COVID-19 and Beyond; Current and Future Steps to Fully Realize this Vision

Jerome A. Osheroff, MD1; Brian S. Alper, MD2; Christopher Tignanelli, MD3; Julia L. Skapik, MD4

1TMIT Consulting, LLC, Naples, FL; 2Computable Publishing LLC, Ipswich, MA; 3Department of Surgery, University of Minnesota, Minneapolis, MN; 4NACHC, Bethesda, MD

Abstract

The ACTS COVID-19 Evidence to Guidance to Action Collaborative is an open learning community supported by AHRQ to help participants respond to the current pandemic. During summer 2021, the Collaborative is describing a compelling ‘art of the possible’ future vision where interoperable and computable healthcare evidence, guidance and data drive whole person care for COVID-19 and other clinical conditions. The future vision includes highly desirable but not yet widely realized patient journeys, as well as needed enhancements to the national / global healthcare knowledge ecosystem to widely realize these outstanding healthcare experiences, processes, and results. This panel will describe this future vision, steps that presenters and others are taking toward realizing it, and additional action needed. After panelist presentations, the second session half of the panel will be interactive discussion with attendees to surface additional opportunities for collaboration and other steps to refine and fully realize the shared future vision.

Introduction

In the U.S. healthcare system, the flow of data, evidence, knowledge, and tools to support critical decisions and actions is highly siloed and fragmented. This leads to preventable problems with care and patient outcomes, unnecessary expenditures, suboptimal patient experience, and overburdened clinicians. Ultimately, the goal – for COVID and beyond – is to achieve the Quintuple Aim to simultaneously enhance patient experience, improve population health, reduce costs, improve clinician experience, and achieve equity in care delivery.

In early 2020, AHRQ launched the ACTS COVID-19 Evidence to Guidance to Action Collaborative in response to the emerging pandemic as a learning community to help participants enhance their efforts to put evolving evidence and guidance for COVID-19 into practice more quickly, efficiently, and effectively. Scores of participants are actively engaged in sharing strategies and tools, synthesizing a shared future vision for a learning health systems approach to the pandemic and beyond, and defining steps to broadly realize that vision. The Collaborative’s capstone deliverable is a concept demonstration illustrating an “art of the possible” journey for a patient with multiple chronic conditions who develops COVID-19, is hospitalized, and subsequently develops “Long COVID” as a complication. This healthcare journey illustrates tools and activities across the continuum of care that are highly desirable and needed to achieve the quintuple aim, but not yet widely realized. The concept demonstration uses this case example as a springboard to describe a knowledge ecosystem that enables efficient flow from evidence to guidance to action to data and back to evidence to produce an effective ‘Learning Health System (LHS)’ cycle (shown below).

Scores of Collaborative participants from government agencies, care delivery organizations, health IT suppliers, specialty societies, standards development organizations, evidence and guidance suppliers, patient advocates, and many other stakeholder groups are working together to synthesize the future vision details and the tools, activities, systems, and infrastructure to broadly realize it. This includes describing steps Collaborative participants and others are taking to drive progress toward this future vision and the needed ecosystem enhancements. The concept demonstration includes a series of deep dives into each area of the LHS/Knowledge Ecosystem Cycle and illustrates how each piece contributes to the patient journey and how all the parts must coordinate to deliver continually improving LHS cycles and the quintuple aim.

Thirty-six organizations representing diverse LHS stakeholder groups have already provided support letters indicating their support for an earlier articulation of the ACTS future vision and Roadmap to broadly achieve it. The letters indicate a desire to collaborate with other stakeholder to realize the vision, including exploring ways to coordinate.
strategies and investments toward that end. AMIA is one of the organizations that provide such a letter⁴, and this proposed late breaking session can help surface specific opportunities that AMIA could pursue to achieve its objectives while accelerating progress toward industry-wide goals.

Proposed Panel

This session builds on the panel session “Using Specific COVID-19 Targets and Patient Care Settings as a Springboard for Driving Global Improvements in the Learning Health System Cycle for COVID-19 and Beyond.” That panel (already on the program) focuses on why the Collaborative was formed, how it has added value to participants’ efforts to address the pandemic and broader LHS needs, and implications for attendees. This proposed ‘Late Breaking Session’ panel focuses on details of the concept demonstration – i.e., a desirable, consensus patient journey for COVID-19 and beyond, and what the needed knowledge ecosystem infrastructure looks like and how to realize it.

Panelists include individuals who have played central roles in the Collaborative and its concept demonstration efforts. The proposed panel’s goal is to showcase the Collaborative’s concept demonstration and through discussion with attendees, identify opportunities to benefit more organizations from the Collaborative’s efforts and accelerate widespread progress toward the future vision. Annual Symposium attendees typically reflect many key LHS stakeholder groups, and this panel is aimed at such participants that have care transformation – for COVID-19 and beyond – as a focus.

Dr. Osheroff, an independent consultant under a contract supporting AHRQ, chairs the Collaborative and will describe the concept demonstration processes and goals, and provide an overview of the future state patient journey. The portion of the patient journey addressing the use of anticoagulation for patients hospitalized with COVID-19 – and other targets - will be addressed by the subsequent panelists.

Dr. Brian Alper, CEO of a company to make science machine interpretable and lead of international volunteer evidence-based medicine standards initiatives, will demonstrate how computable evidence and guidance related to
using anticoagulation in hospitalized patients with COVID-19 supports the patient journey. He will also discuss steps needed to broadly enhance the evidence to guidance part of the LHS cycle for COVID-19 and beyond.

Dr. Christopher Tignanelli, University of Minnesota (UMN), will describe the desired future vision for localizing computable guidance provided by trustworthy sources; putting it into practice via interoperable pathways and clinical decision support interventions; and evaluating results via electronic clinical quality measures. He will describe steps UMN is taking toward realizing this future vision, and broader efforts needed to fully realize the vision.

Dr. Julia Skapik, Medical Director, Informatics for the National Association of Community Health Centers (NACHC) will describe efforts by NACHC and its partners to implement the consensus future patient journey in the healthcare safety net, and enhance the underlying knowledge ecosystem that supports these journeys.

The panel’s second half will be devoted to interactive discussion among panelists and audience members to surface opportunities to further accelerate efforts to achieve this future vision, and leverage steps currently underway toward this vision to enhance attendee efforts. Prompts for this discussion include:

- What areas of the patient journey and future vision resonate?
- Are there opportunities to leverage the vision, steps, and activities described to advance your current work, and to align your efforts to build this future vision and underlying knowledge ecosystem?
- What would that alignment look like and what would be needed to achieve this collaboration?

All participants have agreed to take part in the panel.

**Conclusion**

The COVID-19 pandemic has focused sharp attention on how far healthcare is from optimally putting evidence into practice and navigating other LHS cycle steps - and the urgency and high stakes for closing this gap. The ACTS COVID Collaborative leveraged current efforts and built a concept demo that shows the art of the possible – using a compelling patient journey to demonstrate the future vision of an ecosystem that supports Learning Health System cycles and leverages computable biomedical knowledge to achieve the quintuple aim. This interactive panel will provide attendees with valuable insights and hopefully help identify opportunities to advance their efforts in ways that further accelerate progress toward a shared LHS future vision.

**References**

Reverse clinician burnout trends by exploring clues from global policy variances

Larry Ozeran, MD, FAMIA;1 Jon Patrick, PhD, MAMIA;2 Yalini Senathirajah, PhD;3 William J. Foster, MD, PhD, FRCS, FAMIA;4 Richard Schreiber, MD, FACP, FAMIA, ABPM-CI\textsuperscript{5}

\textsuperscript{1}Clinical Informatics, Inc., Woodland, CA, USA; \textsuperscript{2}Innovative Clinical Information Management Systems (iCIMS); \textsuperscript{3}University of Pittsburgh, Pittsburgh PA USA; \textsuperscript{4}Department of Bioengineering, Temple University, Philadelphia, PA, USA and AltaSciences, Montreal, Québec, Canada; \textsuperscript{5}Penn State Health Holy Spirit Medical Center, Camp Hill, PA, USA

Interactive Panel Programmatic Theme: Clinical Informatics
Tracks: Policy Keywords: Global Health; Legal, Ethical, Social and Regulatory Issues; Usability
Working Group Endorsements: Ethical, Legal, and Social Issues; Clinical Information Systems

Learning Objectives:

- Define opportunities to reduce clinician burnout around the globe
- Learn global policy differences which prevent, reduce, or eliminate root causes of clinician burnout
- Debate panelists’ suggestions for policy and process modification to eliminate root causes of burnout

Abstract

Clinician burnout is a growing problem in the United States and to a much greater extent than in other countries around the world. The Covid pandemic has increased the rate of clinician burnout to the point where there are mounting concerns of clinician shortages in the very near future\textsuperscript{1} and at enormous cost.\textsuperscript{2} In this context, it is increasingly clear that urgent efforts are needed to reverse this dangerous trend. This panel will focus on root causes of clinician burnout, especially on how features of healthcare policies that impact health informatics implementation in regions outside the US seem to reduce or avoid clinician burnout and may model how to reverse the current trends if adopted in the US and in other countries where clinician burnout is prevalent.

Panel Description

Clinician burnout was first described in the United States only 25 years ago.\textsuperscript{3} Its prevalence in the last 10 years has expanded markedly, coincident with adoption and wide dissemination of Electronic Health Records (EHRs). The physical and emotional toll of the Covid pandemic has exacerbated the high baseline level of burnout. Considering the current trajectory of the rate of burnout, it seems likely we will start to see clinician shortages in the not-to-distant future.

There have been many studies, publications, and presentations discussing burnout. Generally, these efforts have focused on how to mitigate the symptoms of burnout rather than eliminate the underlying causes. For example, meditation has been proposed to allow clinicians to calm and refocus, but this does nothing to remove the underlying cause. Although there are studies showing different documentation practices across the globe,\textsuperscript{4} we are not aware of any investigations or discussions that have looked at the impact of international policy variances that might impact the varying rate of burnout in different countries. This panel is the first to our knowledge to consider how the United States and other countries with high rates of burnout might consider implementing major policy changes to reduce rates of burnout to the level seen in most countries, including many where clinician burnout appears to be rare.

Larry Ozeran, MD, FAMIA, will introduce the panel and moderate the discussion. He will guide the audience on a journey from the United States to Canada and Australia to see how differences in health policies lead to differences in informatics solutions that might provide clues to reducing clinician burnout. He will then briefly describe the difference between symptom mitigation and the elimination of root causes of clinician burnout as well as the danger of failing to eliminate the root cases.\textsuperscript{3} With this context he will challenge the audience with questions like:

- If the U.S. reduced the requirement for clinical documentation to match Canada or Australia, how might that change our existing informatics solutions and potentially reduce the rising rate of clinician burnout?
- Can we improve informatics tools and implementation without changing informatics policies?
- How might informatics standardization simplify informatics tools to reduce the transition from one informatics tool to another?
**Jon Patrick**, PhD, MAMIA, will discuss clinician burnout as a function of Clinical Systems Architecture and Interfaces. A study of EHR software in 2008-2010 pointed to major limitations in large scale systems and hardened best-of-breed systems. It identified critical needs for clinical users’ satisfaction with an EHR system: 1. Immediate adaptability, 2. User control of the system design, 3. Native Interoperability, 4. In-built analytics. Early experimental work showed viable technical solutions to these requirements which lead to a commercialization initiative which has resulted in a number of clients. The largest system built with this technology has a been full oncology patient EHR with Tumour Board Management for 11 discrete tumour streams and incorporating direct delivery of data from systems for: Surgery, Pathology (including NLP content extraction), Patient administration, Chemotherapy, Radiology and Radiotherapy. There is no evidence of a serious need for workarounds as only 138 change requests have been serviced at an average time delay of 11 days over 18 months of maintenance. In a clinician conducted assessment at another site it was shown learnability was very high. These cases indicate that clinician control of the design of the EMR results in high levels of satisfaction and so reduces the agents of burnout.

**Yalini Senathirajah**, PhD, will discuss burnout in international perspectives. Other countries may have less of a problem, due to different roles of EHR documentation in the US, where it is inextricably tied to billing and legal liability purposes. She will summarize the association of EHRs and physician burdens and burnout across Anglophone Canada, Europe, and Asia, different top-down and bottom-up approaches and their impact, including modification of US-based software for different countries, vendor control, and designs. She will discuss sociotechnical aspects of EHR development and use, contrasted with consumer app development, and how this makes a difference in EHR usability, possible mechanisms for change, and psychological aspects of EHR use unrelated to usability. Recent European and Asian flexible system initiatives including NHS hybrid model(s) are discussed, and she will cover design and architectural features which reduce cognitive and navigational burdens.

- What pathways exist to test and implement potential solutions arising from clues we find in international work?
- What barriers exist?

**William Foster**, MD, PhD, FRCSC, FAMIA will discuss Canadian physician burnout affecting approximately 30% of physicians, and why implementation of EHRs and electronic tools for clinical practice are substantially lower than in the USA. 11 – 35% of physicians e-prescribe, yet 80 – 92% have digital access to laboratory results. In most provinces, a top-down approach is common: Québec offers free access to Dossier Santé Québec, a database with a secure web-based interface to access patient imaging and laboratory results, and currently filled medications. The majority of outpatient specialists in Québec primarily use paper charts, often with a PACS system, although there is some effort to transition to EHRs (there are few EHRs designed for the French language and the Régie de l’assurance maladie du Québec insurance system). Notes are typically brief. Surveys suggest burnout is less prominent in Canada than in the USA. An increased focus on providing physicians with free informatics tools that provide validated data (from pharmacies, imaging centers, laboratories, perhaps by the use of state-wide registries) without additional effort on the part of the physician may provide direct benefit to physicians and reduce physician burnout. Dr. Foster will discuss other sources of data that can be useful for front-line providers, and ask what informatics approaches might mitigate physician burnout.

- Is it possible to provide free informatics tools that offer physicians validated data?
- Can we find reliable, clinically relevant data without physician review prior to entry into the medical record?

**Richard Schreiber**, MD, FACP, FAMIA will discuss the profound deterioration in clinician satisfaction in the United States during the past 30 years. Many ascribe this to the introduction of the EHR resulting in abrupt retirement of some clinicians, but other factors contribute. Dr. Schreiber will elaborate on these questions:

- Can we improve usability, and eliminate workflows and tasks to reduce documentation burden?
- What policy and regulatory changes can effect real change in documentation burden?
- What is the impact of two apparently contradictory documentation needs: how to capture sufficient medical information (and yet not cause note bloat) while providing sufficient association of that information for billing?
- What lessons learned has the Covid pandemic provided that may actually reduce burnout?
- How can we bring the pleasure of clinical work back to the bedside to reduce clinician burnout?

**Participants (All panelists have agreed to participate)**

Dr. Ozeran has spent more than 20 years in information technology and software development, clinical medicine, healthcare leadership, and health policy. His efforts have ranged from the development and interpretation of public policy to helping startups develop market strategy to supporting healthcare organizations. Dr. Ozeran has responded to proposed federal regulations, written to many elected officials, and testified before committees of the California Senate and Assembly. He was an invited speaker before the Congressional Budget Office. He has twice served on the AMIA Public Policy Committee. He is focused on strategic initiatives, policy development, and innovation supporting healthcare entities, government agencies, and startup companies.
Dr. Patrick In 2005 Dr. Patrick won Australia’s national Eureka Science prize for his work in natural language processing. Subsequently he has conducted extensive research on the use of language technology in Intensive Care, Pathology and Radiology departments, and in information systems research in emergency medicine and oncology. In 2012 he left the University of Sydney, where he held both the Chairs of Information System and Language Technology, to pursue his interests in commercializing his Health IT and NLP ideas. He is the CEO for the companies Health Language Analytics (HLA) and its USA subsidiary Health Language Analytics Global, and Innovative Clinical Information Management Systems (iCIMS). His NLP companies hold contracts for the 3 California Cancer Registries and the Centers for Disease Control, and the University of California and hospitals and authorities in Australia. His clinical systems company has built over 30 different applications particularly in the arena of cancer care and tumour board solutions. Most recently the work of the two companies at the Sydney Adventist Hospital has been highlighted at https://www.pulseimagazine.com.au/australian-ehealth/4980-the-san-rolls-out-icims-cancer-information-management-system-for-mds.

Dr. Senathirajah conducts AHRQ-funded studies of EHR design and usability, including studies of interruptions, cognitive load imposed by different EHR UI designs, and a novel ‘composable’ approach which permits the nonprogrammer clinician end-user to design and control their own interfaces by drag/drop, assembling relevant data on the same screens and marking, visualizing and sharing it in different ways. The aim is to reduce display fragmentation and clinical cognitive, navigational, and time burdens. She is Associate Professor in the department of biomedical informatics at University of Pittsburgh School of Medicine.

Dr. Foster is an American-trained Ophthalmologist and physician-scientist with experience practicing in the US and in Canada. He earned his undergraduate degree in Physics from Caltech, a Master’s and PhD in Physics from Harvard University, and his MD from Duke. He is a Diplomate of the American Board of Ophthalmology and the American Board of Preventive Medicine in Clinical Informatics as well as a Fellow of the Royal College of Surgeons of Canada. He is a Fellow of both the Association for Research in Vision and Ophthalmology and AMIA. He has extensive experience with implementation and clinical decision support with electronic medical records and, for example, is a Physician Builder with numerous Epic certifications.

Dr. Schreiber is Associate Chief Medical Informatics Officer, Penn State Health Holy Spirit Medical Center, Camp Hill, PA; Professor of Medicine, Geisinger Commonwealth School of Medicine; and Chair of the Clinical Information Systems Working Group. His research focuses on clinical decision support, alert fatigue and alert burden reduction, transitions from one EHR to another, drug-drug interaction alerting, documentation improvement, problem list curation, ethical issues in informatics, and laboratory nomenclature. Recent publications include risks of harm of mHealth applications, and a perspective on reduction of clinician burnout.

References
Adaptable Patient-facing and Clinical Decision Support Systems: The Next Frontier

Mustafa Ozkaynak, PhD¹ (Co-organizer), Karen D. Lopez, PhD, MPH, RN² (Co-organizer), Adam Wright, PhD³, Andrew Boyd, MD⁴, Blackford Middleton, MD, MPH, MSc⁵ (Moderator)

¹University of Colorado | Anschutz Medical Campus, Aurora, CO, USA; ²University of Iowa, Iowa City, IA, USA; ³Vanderbilt University Medical Center, Nashville, TN, USA; ⁴University of Illinois at Chicago, Chicago, IL, USA; ⁵Apervita Inc., Chicago, IL, USA

Abstract
Clinical decision support (CDS) systems can demonstrate their full potential if they are context-sensitive and can tailor their interaction with the user. Adaptable CDS aims to achieve this goal by accounting for various context and user characteristics and therefore avoid the many undesired outcomes of current CDS systems. To this end, the panel will begin with a discussion by each panelist of their experience with adaptable CDS in various health care settings with decision support that targets both clinicians and patients. This will be followed with an interactive session, in which the audience and panelists will prioritize related challenges (generic, or clinical setting specific) in development of adaptable CDS. Finally, we will discuss known effective and potentially new strategies to address these challenges.

Intended Audience
Attendees of any level who are interested in: 1) the next generation of clinical decision support (CDS) that aims to reduce alert fatigue and cognitive workload and 2) improving the efficacy of these systems. This interdisciplinary panel considers CDS broadly to include both adaptable patient- and clinician-facing decision support (CDS) systems. The panelist will present these next generation CDS systems, which are adaptable to user or context characteristics, from the perspectives of industry, practice, and research. Attendees will benefit from this session by understanding: 1) the adaptive mechanisms that make technology successful in various industries, 2) novel approaches to study adaptability, and 3) ways to integrate the current research into the design, implementation, and evaluation of CDS.

Introduction of the Topic
CDS should evolve rapidly, given the exponential growth in health data and evidence, and the growing capacities to store and analyze data.¹ Future CDS can address this necessity by being adaptable. Adaptable CDS is defined as CDS with the ability to change its interaction with the user, based on programmed sensitivities to various issues and characteristics, in order to provide optimal decision support. Adaptability to such macro factors (e.g. new evidence, changes in policies) and micro factors (e.g. user or situational characteristics) is a necessary capability for CDS to provide the best support to clinicians who are making decisions that directly affect patient outcomes.

An important criticism of CDS is that it is not sensitive to the user or the context.²⁻⁵ Alarm fatigue is the most well-known issue of decontextualized CDS. However, dealing with alarms may cause additional problems, such as excessive cognitive workload. Adaptivity can make CDS context-sensitive by accounting for various macro or micro level factors, and present clinicians with opportunities to improve care through individualized interaction with CDS technology and patient-centeredness.¹ For example, CDS can adapt to user characteristics, such as the user’s experience or fatigue level, or context characteristics (e.g. busyness of clinics, up-to-date evidence).

Although adaptable CDS is novel in health care, adaptable technologies have been studied in other domains such as defense systems and transportation. Existing frameworks and approaches can provide a head start in designing adaptable CDS.

Aim of Discussion
This session will highlight the opportunities that adaptable CDS technologies can bring to clinicians and patients in clinical practice, self-management, and wellness in daily living. The panel will discuss the challenges and facilitators of the design, implementation, and evaluation of adaptable CDS. We will engage in an exchange of ideas and experiences from both the expert panel and the audience. The panel will focus on novel features (i.e. adaptability) that
can be a standard feature for the next generation of CDS. This panel will link interested individuals with one another, and mutually learn about cutting-edge work in this area.

Table 1. Scheduled topics and panelists.

<table>
<thead>
<tr>
<th>Time</th>
<th>Topic</th>
<th>Speaker</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 min</td>
<td>Introductions, framing of the discussion</td>
<td>Middleton</td>
</tr>
<tr>
<td>12 min</td>
<td>Adaptable Automation and Fatigue</td>
<td>Ozkaynak</td>
</tr>
<tr>
<td>12 min</td>
<td>Adapting based on cognitive characteristics</td>
<td>Lopez</td>
</tr>
<tr>
<td>12 min</td>
<td>Incorporating user feedback and performance data to support CDS adaptability</td>
<td>Wright</td>
</tr>
<tr>
<td>12 min</td>
<td>Adapting CDS based on patient context, motivation and priorities</td>
<td>Boyd</td>
</tr>
<tr>
<td>37 min</td>
<td>Discussion, Q&amp;A with audience</td>
<td>All</td>
</tr>
</tbody>
</table>

Timeliness of the Topic

CDS systems show promise in improving patient and organizational outcomes. Their full potential can be realized only when they become sensitive to the relevant characteristics of users and the context in which CDS is used and decisions are made. Because of its interactive nature, the panel will serve as a forum in which members of the audience and panelists can collaboratively advance the dialogue of informatics in this novel area. The sharing of experiences stimulated by this panel will serve as a foundation for generating and prioritizing future design and research initiatives in CDS at various health care delivery settings.

Contribution of Each Panelist

Mustafa Ozkaynak: Dr. Ozkaynak will discuss adaptive (or adaptable) automation literature that can provide insights into developing next generation of CDS. He will also discuss being adaptive to clinicians’ fatigue, a significant barrier to the clinician successfully adapting to CDS. Relevant literature will be reviewed, as well as findings from a study that aimed to develop CDS as a part of antimicrobial stewardship program in emergency departments.

Adaptive automation, an emerging area within human factors engineering, is an approach to technology design where tasks are dynamically allocated between user and technology. One main system parameter of adaptive automation is the level of automation, which depends on the amount and type of work allocated to user and computer.6

Decades of adaptive automation literature suggest that CDS systems that can adapt to contemporaneous variations in clinician’s fatigue have the potential to intercept fatigue-induced human errors and preclude potential adverse events.5 9 Adaptive modifications to CDS for fatigue have many benefits including: (1) lowering trigger rates, thus reducing alert fatigue; (2) providing an interaction opportunity that matches the user’s current capability; (3) compensating for suboptimal working conditions; and (4) reducing information overload to increase the effectiveness of decision-making. Despite the potential benefits of adaptable CDS, it must be designed and implemented deliberately in order to avoid the possible negative effects of having multiple behaviors of the same CDS during the same shift.

Our recent study,3 highlights the significance of fatigue in EDs in making antibiotic decisions. Adaptability can be a feasible and effective solution to incorporate clinician and organizational factors into CDS design and implementation.

Karen Dunn Lopez: Dr. Dunn Lopez will share insights from her multidisciplinary NIH-funded research team that applied data mining algorithms from nursing care documentation data in electronic health records (EHR) from four hospital systems.10 The analysis provided the foundation for alerts that were transformed into multiple CDS formats, which were then tested using simulation with 60 practicing nurses.11 The findings suggested differences in comprehension and efficiency of use among the participants. We hypothesize that differences may be associated with cognitive characteristics (graph literacy and numeracy) and format preference.12,13 We tested this hypothesis in our follow-up randomized controlled trial using simulation methods with a fully powered national sample of nurses. We tested our adaptable CDS web app remotely from nurses’ homes using video conferencing software. Our methods provide a feasible foundation for widespread national testing of adaptable CDS before use in practice.

Adam Wright: Dr. Wright will share synthesized findings from multiple projects that provided insights about challenges of and effective strategies for gathering and incorporating user (e.g., clinicians) feedback and performance (e.g., specificity and sensitivity) – data that can lead to CDS that is adaptive to situation and local context. He will specifically highlight design and implementation implications of these insights.
Andrew Boyd: Dr. Boyd will discuss insights learned through a multi-year study of tailoring patient’s discharge instructions of patients with heart failure. This tailoring and simplification is important as patients receive lengthy documents at the time of hospital discharge that include complex terms and, occasionally, conflicting information provided by physicians and nurses. This has become more routine with the OpenNotes initiative and is particularly challenging for patients with heart failure in settings that serve patients in the low socioeconomic group. These patients have unique challenges in comprehending the information, which intends to support effective self-management at home and help them avoid hospitalization. Our multidisciplinary NIH-funded research team addressed these challenges through the design of a novel mHealth for post-hospitalization heart failure self-management. Their design used natural language generation to merge physician and nursing care summaries into one information source, which was tailored to patients’ context and motivation, including their health literacy and engagement level.14

Interactive Session

After the introductory presentations, the panel and audience will engage in a facilitator-moderated discussion of (1) important design, implementation, and evaluation considerations for adaptable CDS; (2) information needs to make CDS more adaptable; (3) potential unintended consequences of adaptable CDS and mitigation strategies; and (4) ways that the AMIA community can advance the field of adaptive CDS. Prioritized lists of issues and strategies will be developed for research, practice, and design. All panelists have agreed to take part on this panel.

References

To share or not to share: Exploring the ethical implications of sharing personal health data with patients and informal caregivers

Meghan Reading Turchioe, PhD, MPH, RN; Sabrina Mangal, PhD, RN; Marianne Sharko, MD, MS; Natalie Benda, PhD; Ruth Masterson Creber, PhD, MSc, RN

Weill Cornell Medicine, Dept. of Population Health Sciences, Div. of Health Informatics, New York, NY

Introduction

National conversations are taking place about ethical practices of ownership, sharing, and privacy of electronic personal health data as it proliferates in healthcare. To date, a major focus of these conversations has been data sharing between healthcare organizations or companies. There has been significantly less attention paid to ethical practices surrounding organizations returning personal health data back to patients or informal caregivers. While there are at least legal standards concerning patients’ personal health information that compel healthcare organizations and researchers to take precautions, such as the Health Insurance Portability and Accountability Act (HIPAA), there are no explicit ethical obligations for them to return personal health data to patients or their informal caregivers. This creates moral ambiguity regarding the return of personal health data; withholding the data may cause harm, but returning it without adequate consideration of nuances or comprehension could also cause harm.

One example includes the collection and return of patient-reported outcomes (PROs), defined as health outcomes directly reported by the patient, back to patients. PROs are widely used in clinical trials and increasingly collected by healthcare organizations nationally for use in clinical care. In cancer care, use of PROs by health professionals has been demonstrated to improve survivorship in several clinical trials, and thus is becoming part of routine clinical practice. Despite the acknowledged value of PROs, ethical obligations surrounding whether PRO data should be returned to patients are in nascent stages. In general, patients report that viewing their PROs is highly informative because they provide an indicator of current health status (for example, symptoms, functioning, and quality of life). When tracked over time, PROs provide insight into changes in health status, and can also empower patients. However, PROs are rarely returned to participants. At the same time, there are concerns about the potential to cause harm when PROs are not returned with adequate explanation, context, or guidance for patients to correctly interpret the data and decide whether and how to act based upon the data.

Another example is parental access to pediatric health portal accounts. There is concern for adequate privacy protection related to sensitive health information released through patient portals. With increased use of OpenNotes and other forms of health data accessed through patient portals, inadvertent breaches of privacy are possible. This is concerning for vulnerable patient populations, such as adolescents whose parents may have access to patient portal information. Considerations regarding the type of information to release in a patient portal must take into account potential harm to patients and their families. Clinically, adolescents are likely to have new and potentially sensitive medical needs, such as sexual and reproductive healthcare and mental health issues. Mental health visits increase during adolescence and can include screening surveys for reported mental health symptoms, such as depression and anxiety. Studies have found that adolescents need privacy and confidentiality when deciding whether or not to receive medical care and how much sensitive information to share with their providers. Inconsistent privacy protection can erode trust in healthcare providers. Until professional standards are developed and technology reliably protects patient privacy, the risk of inadvertent privacy breaches remains.

Additional ethical questions arise when considering other types of informal caregivers beyond parents. For example, informal caregivers play an important role for cancer patients, reportedly spending an average of 32.8 hours on care-related tasks per week. Patient portals and portal secure messaging provides a means for patients and their informal caregivers to communicate with healthcare professionals. One problem, however, is that portal secure messages coming from a patient account, as opposed to a formally designated proxy account, may actually be written by someone other than the patient. Promoting the appropriate use of proxy patient portal accounts will help understand the needs of informal caregivers, differentiate patient and caregiver needs, and ensure the patient has agency in granting proxy access.

In this panel, we will present primary data from multiple studies exploring these ethical questions from the perspective of multiple key stakeholders, including patients, informal caregivers (including parents), and healthcare professionals. These include qualitative interviews with stakeholders and surveys with a large, representative sample of U.S. adults as part of an NIH bioethics supplement grant (R00NR016275-05S1). In addition, we will discuss opportunities to bring diverse perspectives into the public discourse on policies and governance of personal health data, ownership, privacy, and use.
Each ethical question will be explored through the lens of an ethical framework developed by MITRE, “An Ethical Framework for the Use of Consumer-Generated Data in Healthcare.” The framework is intended to guide healthcare organizations and professionals to consider salient ethical concepts when creating policies surrounding personal health data access. In this panel, we will discuss each of the three example scenarios described above in the context of core values which must be protected according to the framework, including distributive justice, individual self-determination, privacy, and trustworthiness. We will also explore policies for each scenario that adhere to guiding ethical principles, including fairness, individual and population health, empowerment, and transparency.

**General Description of Panel and Objectives**

Ethical quandaries surrounding the return of personal health data to patients and informal caregivers merit serious consideration given the potential for both sharing and withholding data to cause harm. In this panel, we will explore and synthesize the ethical principles and values that surround the return of personal health information from the perspectives of multiple stakeholders within the clinical enterprise who regularly engage with electronic personal health data. In addition to presenting primary data from multiple studies with key stakeholders, we will also discuss potential policies that may best uphold the core values and guiding ethical principles within the rapidly changing digital health landscape. By the end of this panel, participants will be able to:

- Identify ethical principles and values relevant to the collection, use, and sharing of personal health data.
- Discuss potential benefits and harms of returning personal health data to patients and informal caregivers.
- Apply the MITRE framework to generate ethical guidelines for the return of personal health data to patients and informal caregivers in specific contexts.

**Intended audience**

As digital tools are increasingly used to generate and share personal health data across academic and industry research, questions of appropriate recipients of the data, including patients themselves, are highly pertinent. Therefore, this topic is relevant to informatics researchers, administrators, and clinicians in industry, academia, and healthcare, particularly those who are generating policies and guidelines surrounding personal health data access.

**Interactivity**

This panel will incorporate multiple features to facilitate engagement and interaction with the audience, including live audience polling using Poll Everywhere and a moderated discussion. Differing viewpoints will be encouraged to stimulate rich dialogue surrounding each of the example scenarios. Example questions include:

- How should researchers and healthcare organizations weigh the potential benefits against the potential harms of sharing data with various recipients (i.e., patients, different types of informal caregivers)?
- How can the MITRE framework be operationalized to create policies that protect patients from harm?

Example policies will be constructed through dialogue from the audience.

**The Panel**

**Dr. Meghan Reading Turchioe** is a nursing informatics researcher whose work involves developing and implementing consumer health technology to improve health outcomes and quality of life. Specifically, she studies ways to develop and implement consumer technologies including mobile applications, wearables, and telehealth to support optimal patient self-management, shared decision-making, and prevention for chronic conditions. Dr. Reading Turchioe will describe the legal and ethical landscape surrounding access to personal health data and introduce the MITRE framework, including salient core values and guiding principles to be considered.

**Dr. Sabrina Mangal's** research focuses on patient engagement and communication strategies (e.g., visualizations) to
enhance patients’ comprehension of their health information. She is closely involved in studies that explore the potential for enhanced engagement in care. Dr. Mangal will discuss the bioethical implications of researchers returning collected health information to participants, and public levels of trust in research based on sharing collected health information to multiple stakeholders.

Dr. Marianne Sharko is a pediatrician who received her master’s degree in Health Informatics from the Population Health Sciences Department at Weill Cornell Medicine and has completed a Preventive Medicine Residency. She is currently working in a pediatric obesity clinic and conducting research through the Empire Clinical Research Investigator Program. Her research interests include protection of patient privacy in the electronic health record and patient portals, particularly for adolescent populations. Dr. Sharko will discuss the unique patient privacy needs of the adolescent patients in relation to the patient portal and parental proxy access.

Dr. Natalie (Nat) Benda is an expert in human factors engineering/systems safety, and her work involves advancing health equity and inclusivity through technology-based interventions. She will introduce a project about using natural language processing to automatically detect patient portal secure messages in a cancer population sent from the patient account by someone other than the patient (e.g. an informal caregiver), a phenomena referred to as “hidden proxies”. In this work, the team is also analyzing demographic differences in patient accounts containing a larger proportion of hidden proxy messages. Dr. Benda will present these results and discuss implications for proxy data sharing in an oncology setting.

Dr Ruth Masterson Creber leads an interdisciplinary program of research focusing on using technology and remote monitoring data to improve patient centered outcomes among individuals with cardiovascular disease. Dr. Masterson Creber is the principal investigator of NIH-funded grants that support the data collection and return of patient reported outcomes (R00NR016275, R00NR016275-05S1, R01HL152021) and a PCORI grant evaluating a clinical informatics intervention, “Using Mobile Integrated Health and Telehealth to Support Transitions of Care among Heart Failure Patient.”

Panel Organizer Statement: All participants have agreed to participate in this panel.

Conflicts of Interest: Dr. Reading Turchioe is affiliated with Iris OB Health Inc. and has equity ownership. 

Acknowledgements: The described research was supported by the National Institute of Nursing Research (4R00NR016275-05S1) and Memorial Sloan Kettering Cancer Center’s Population Sciences Research Program (PI; L Diamond and P Stetson). Dr. Sharko is supported by the NYS Department of Health Empire Clinical Research Investigator Program. The content is solely the responsibility of the authors and does not represent the views of these agencies.

References

Perspectives on Developing and Implementing Shareable, Interoperable Clinical Decision Support for Chronic Pain: The CDS4CPM Project

Joshua E. Richardson PhD, MS, MLIS¹ (Organizer); Cheng-Kai Kao, MD²; Laura Haak Marcial, PhD¹; Bryn Rhodes³; Asli Weitkamp, PhD⁴

¹RTI International, Research Triangle Park, NC; ²University of Chicago, Chicago, IL; ³Alphora, Orem, UT; ⁴Vanderbilt University Medical Center, Nashville, TN

Abstract

Open-source efforts that utilize emerging standards for clinical decision support (CDS) may enable healthcare organizations to share and implement evidence from external stakeholders including guideline clearinghouses and researchers. This panel provides perspectives from an Agency for Healthcare Research and Quality-sponsored project, CDS for Chronic Pain Management (CDS4CPM), that are working with publicly-available standards-based, interoperable CDS for chronic pain management and shared decision-making in real-world settings. The panelists will provide their unique perspectives and experiences as developers and implementers of these CDS knowledge artifacts and detail how this work can inform future efforts to develop and implement similar CDS interventions. The CDS4CPM speakers will highlight the technical and organizational challenges they encountered over the course of the project and provide their insights for audience members interested in the challenges and opportunities for leveraging new open-source CDS tools and standards, particularly in chronic pain management.

Significance

Chronic pain affects as many as one in ten Americans and significantly disrupts “work, social, and/or self-care activities.”¹ Furthermore, the related use of opioids is linked to co-morbidities and mortality that some have described as “twin crises” in healthcare,² and the annual costs of chronic pain exceed those from heart disease, diabetes, or cancer ($560 billion to $630 billion).³ The COVID-19 pandemic further complicates chronic pain and opioid management.⁴ The significant rates of morbidity, mortality, increased costs and adverse effects on quality of life make chronic pain a significant problem in healthcare and is therefore an important area for informatics research. Given this context, public access to standards-based and shareable clinical decision support (CDS) in the future may help providers and healthcare organizations deliver timely and evidence-based patient-centered care.

Clinical decision support development tools and strategies are undergoing rapid changes due to the emergence of shareable interoperable solutions like those deposited in the Agency for Healthcare and Research Quality (AHRQ)-sponsored CDS Connect Repository of available CDS artifacts. These artifacts leverage Health Level 7 (HL7) standards for CDS including Clinical Practice Guidelines on FHIR (CPGonFHIR), CDS Hooks, Clinical Quality Language (CQL), and SMART on FHIR. AHRQ contracted RTI International (RTI) to address two broad aims: 1) to advance evidence into practice through CDS; and 2) to make CDS more shareable, interoperable, and publicly-available. RTI in turn partnered with two academic medical centers, Vanderbilt University Medical Center (VUMC) and the University of Chicago Medicine (UCM), a software development company (Alphora), and consultants to develop and implement patient-facing and provider-facing CDS artifacts intended to promote shared decision-making around chronic pain management. The team named the project CDS for Chronic Pain Management or CDS4CPM.

After a year of development, the team implemented all components of the CDS4CPM system. This includes a patient-facing CDS artifact called MyPAIN (My Pain Assessment and Information Needs) as a SMART on FHIR application that launches from a patient portal (MyChart, Epic). MyPAIN is used to collect patient reported data including pain location and type, PROMIS measures on pain intensity and interference with daily functioning, and patient goals and values. MyPAIN also provides education materials related to chronic pain management. The data collected via MyPAIN are then stored in the electronic health record (EHR) and retrieved by a provider-facing SMART on FHIR application, PainManager, which is available within the EHR. PainManager displays MyPAIN data as well as relevant diagnoses, medications (including PDMP data when available), and urine drug screens. PainManager offers both passive and active CDS alerts based on the CDC Opioid Guideline.⁵ PainManager also provides individual medication
level morphine medication equivalent (MME) and total MME/day calculations using an embedded calculator, also based on the CDC Opioid Guideline.

During development and implementation, project members routinely addressed issues that were not only technological in nature, but also challenges related to integration into the local environment, including navigating internal development processes, directing staff, and engaging patients and providers. Examples of these included concurrent testing of the CQL-based MME calculator (technical); determining a means to reconcile and access data from a state PDMP based on policy and legal limits (organizational); and weighing strategies for CDS maintenance to support the CDS beyond the project period (financial). Developing and implementing standards-based, interoperable, and publicly available CDS is socio-technically complex, not simply “plug and play” for even sophisticated implementation sites. Panel members will discuss these and other challenges along with solutions encountered over the course of the CDS4CPM project, with each panelist sharing their unique perspective and lessons learned.

Description of the Panel

<table>
<thead>
<tr>
<th>Time</th>
<th>Speaker</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>10’</td>
<td>Richardson (moderator)</td>
<td>Introduces the panel and provides an overview of the CDS4CPM project, aims, and goals—and then leads discussion during the audience Q&amp;A</td>
</tr>
<tr>
<td>12’</td>
<td>Marcial</td>
<td>Describes the technical approach to developing two CDS artifacts using standards-based approaches including the collection of patient-reported outcomes data</td>
</tr>
<tr>
<td>12’</td>
<td>Rhodes</td>
<td>Discusses the challenges and solutions for leveraging Clinical Quality Language for CDS logic, utilizing SMART on FHIR APIs, and developing a standards-based IG</td>
</tr>
<tr>
<td>12’</td>
<td>Weitkamp</td>
<td>Discusses VUMC’s project challenges, achievements, and lessons learned</td>
</tr>
<tr>
<td>12’</td>
<td>Kao</td>
<td>Discusses UCM’s project challenges, achievements, and lessons learned</td>
</tr>
<tr>
<td>30’</td>
<td>All</td>
<td>Discussion, Q&amp;A with audience</td>
</tr>
</tbody>
</table>

Learning Objectives

1. Describe technical challenges and opportunities for healthcare organizations developing a single system with publicly-available, standards-based, and interoperable CDS that addresses chronic pain management.
2. Identify organizational strategies that future organizations may encounter—such as staffing and collaborating—when jointly implementing a single system with publicly-available, standards-based, and interoperable CDS.
3. Develop internal and external organizational strategies that account for socio-technical issues of collaborative development and implementation of a single system with publicly-available, standards-based, and interoperable CDS.

Individual Speaker Contributions

Joshua Richardson, PhD, MS, MLIS – Dr. Richardson is a health informatician at RTI International who will serve as the panel moderator and introduce the panel participants. In addition, he will describe the aims of the project to develop, implement, and evaluate the lessons learned for disseminating publicly-accessible, standards-based knowledge artifacts for chronic pain management within primary care clinics. Dr. Richardson will describe the partnering structure with the developer, Alphora, two academic medical centers, VUMC and UCM, and consultants.

Laura Haak Marcial, PhD – Dr. Marcial is a health informatician at RTI International who will discuss specific design choices, technical challenges and opportunities for enhancements associated with patient- and clinician-facing interoperable CDS for chronic pain.

Bryn Rhodes – Bryn Rhodes will discuss development and implementation of the standards-based CDS artifacts, MyPAIN, PainManager, and the related CDS4CPM implementation guide. He will provide an overview of the architectural approach, a review of the standards and content that were selected and used, as well as gaps identified during the project, and some novel approaches to addressing those gaps within the implementation.
Asli Weitkamp, PhD – Dr. Weitkamp is an Associate Professor in the Department of Biomedical Informatics and the Director of Knowledge Engineering at Vanderbilt University Medical Center. She is the site co-Principal Investigator leading the implementation and integration of the CDS for chronic pain management. She will share the experiences, challenges, and lessons learned about integrating standards based CDS into the EHR. She will discuss the strategies for supporting architectural solutions to facilitate data integration between the EHR and interoperable CDS.

Cheng-Kai Kao, MD – Dr. Kao is the Associate Chief Medical Information Officer at University of Chicago Medicine and its site co-Principal Investigator. He will share the experiences, challenges, and lessons learned about implementing a newly developed CDS solution into their EHR, the strategies they used to overcome barriers to adoption, and the considerations and next steps after the pilot phase of the project.

Expected Discussion and Discussion Questions
We expect the audience will want to engage the panelists in discussions on their approaches to leveraging recent open-source standards for developing and implementing patient- and provider-facing CDS. We expect the audience will benefit from learning how our organizations had to adapt staffing and internal processes to operationalize the CDS into clinical workflows. We anticipate these topics to lead to areas for future research and development that could better scale publicly available, standards-based, and interoperable CDS, considerations around implementing approaches to standards-based CDS, and approaches to patient-facing CDS systems. The discussion questions include:

1. How can future projects balance the availability and use of standards-based and proprietary APIs for publicly-available, standards-based, and interoperable CDS?
2. What advances in semantic vocabularies, value sets, and patient-reported outcome measures could benefit the future use of publicly-available, standards-based, and interoperable CDS?
3. How can organizations plan, staff, and budget for working with this kind of CDS?

Urgent Topics for Intended Audiences
This panel addresses pressing issues around developing, implementing, and managing the challenges they encountered with managing the sociotechnical and management challenges publicly-available, standards-based, and interoperable CDS into clinical workflows that address both patients’ and providers’ needs.

- **CMIOs and CNIOs**: those responsible for implementing CDS in healthcare organizations;
- **EHR Implementers**: implementation staff responsible for implementation and effective use of EHR tools, and system optimization, including patient-facing tools;
- **CDS Systems Developers**: development staff building clinical decision support tools and services for providers and patients;

Attestation
The panel moderator has assurances from all participants that they will be available to participate at AMIA 2021.

References
Practical Approaches to Telehealth Equity

Jorge A. Rodriguez, MD¹, Amy Sheon MPH PhD²-⁴, Sarah C. Nosal MD FAAFP⁵,⁶, Courtney Lyles PhD⁷, Jessica S Ancker MPH PhD⁸

¹Division of General Internal Medicine and Primary Care, Brigham and Women’s Hospital and Harvard Medical School, Boston, MA; ²Departments of Bioethics, Family Medicine and Population & Quantitative Health Sciences, Case Western Reserve University School of Medicine Cleveland, OH, ³Public Health Innovators LLC, Cleveland, OH, ⁴Telehealth Equity Coalition, Cleveland, OH, ⁵The Institute for Family Health, Bronx, NY, ⁶Mount Sinai Department of Family Medicine & Community Health, Bronx, NY, ⁷University of California San Francisco, Departments of Medicine and Epidemiology & Biostatistics, Center for Vulnerable Populations, San Francisco, CA, ⁸Department of Biomedical Informatics, Vanderbilt University Medical Center Nashville, TN

Abstract

During the COVID-19 pandemic, social distancing requirements prompted health systems, clinicians, and patients to rapidly switch to telehealth (video and telephone visits) as alternatives to in-person visits.¹ This shift was further supported by policy changes—for example, CMS offered payment parity across visit type.² Early evidence on telehealth use has demonstrated disparities in use, especially as it relates to video visits, among racial/ethnic minorities as well as limited English proficient, low income and low literacy patients (Figure).³ These disparities are driven by factors including a lack of technology access (broadband and device), limited technology literacy, and system level barriers. Thus, there is a need to understand and share best practices for equitable implementation of telehealth. Our panel presents diverse experiences and best practices for telehealth equity.

Learning Objectives

After attending our panel, participants will be able to:

1. Understand the current multilevel drivers of telehealth disparities, including technology access, technology literacy, and workflow considerations.
2. Formulate a diverse set of approaches to improve telehealth equity, including technology screening and digital health navigators.

*Excludes ambiguous encounters.

***p < .001.

Figure 1. Telemedicine Visit Type by Patient Characteristics at Mass General Brigham³
Importance of Topic
The confluence of telehealth as a new care tool, telehealth access disparities, and worse COVID-19 outcomes among underserved communities positions telehealth equity as a critical component of health equity.

General Description
Our panel will present multilevel approaches to telehealth equity from diverse perspectives, including researchers, clinicians, and administrators.

<table>
<thead>
<tr>
<th>Panelist</th>
<th>Presentation Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Panel moderator: Jessica S Ancker</td>
<td>Dr. Ancker has studied a number of issues in health literacy and technology adoption, including ways to reduce socioeconomic disparities in patient portal adoption. She is an associate professor at Vanderbilt University Medical Center in the Department of Biomedical Informatics. Dr. Ancker currently leads a PCORI grant to study the impact of the telehealth transition in primary care prompted by the COVID-19 pandemic.</td>
</tr>
<tr>
<td>Amy Sheon</td>
<td>Dr. Sheon will review several approaches to telehealth equity including: Policy measures being pursued by The Telehealth Equity Coalition; Digital health equity metrics developed for health care systems; Mapping to reveal digital redlining; Implementing universal patient screening for digital health readiness; Establishing community referral partnerships with digital inclusion organizations; and using Digital Health Navigators for screening, referral, and coaching to address individual patient needs for devices, connectivity and skills.</td>
</tr>
<tr>
<td>Courtney Lyles</td>
<td>Dr. Lyles will cover ongoing approaches and best practices for telehealth implementation from safety net delivery systems in California, including a) broad approaches outlined in our telehealth for equity toolkit and b) more specific findings related to language access for patients with limited English proficiency and workflows for team-based primary care.</td>
</tr>
<tr>
<td>Sarah Nosal</td>
<td>Dr. Nosal will walk through the rapid roll out of telehealth in the very early pandemic in a network of Federally Qualified Health Centers located in Brooklyn, Manhattan, Bronx, and Mid-Hudson region. She will discuss transitioning to telehealth in the earliest pandemic hit zip codes in our country in an organization that had had no large-scale telehealth ambitions in its near future. She will highlight choices made to assure patient access and equity in care and transitions made to make telehealth sustainable.</td>
</tr>
<tr>
<td>Jorge A. Rodriguez</td>
<td>Dr. Rodriguez will provide a clinical perspective into the facilitators and barriers for implementing equitable telehealth in primary care and in care transitions following hospital discharge. He will also discuss how privacy, and trust influence telehealth among vulnerable populations.</td>
</tr>
</tbody>
</table>
Discussion Questions
1. What are optimal telehealth outcomes and use cases that we should look for? (e.g., generally video visits are preferred but there are some circumstances where audio and asynchronous might be optimal)
2. How do we know what telehealth use patterns are? Where are data gaps?
3. Which populations are less likely to use video visits as compared to audio or in person?
4. What are factors associated with lower telehealth use, particularly video visits?
5. How can telehealth address disparities for specific diseases?
6. Beyond technology access, how does technology literacy contribute to telehealth disparities?
7. How do different institutions approach telehealth equity?
8. How can healthcare systems screen patients for their technology needs?
9. What are digital health navigators and how can they address telehealth equity?
10. What resources are available to patients who lack the basic technology needs?
11. What are especially considerations for implementing telehealth with patients with limited English proficiency?
12. What are specific privacy and security considerations for implementing telehealth among vulnerable populations?
13. How can public policy influence telehealth disparities?

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

References
AMIA 2021 Session Proposal Didactic Panel on

How the COVID-19 Pandemic Accelerated AI Technology Development and Adoption in Healthcare: Lessons Learned

Panel Organizers:
- Michal Rosen-Zvi, PhD, IBM Research, Haifa, HA, Israel & Vising Professor, Faculty of Medicine, Hebrew University
- Anil Jain, MD, FACP, Independent healthcare IT advisor, part-time clinician, Department of Internal Medicine, Cleveland Clinic, Cleveland, Ohio
- Eileen Koski, M.Phil, FAMIA, Center for Computational Health, IBM Research, Yorktown Heights, NY

Panel Participants:
- Michal Rosen-Zvi, PhD, IBM Research, Haifa, HA, Israel & Vising Professor, Faculty of Medicine, Hebrew University (Moderator)
- Anil Jain, MD, FACP, Independent healthcare IT advisor, part-time clinician, Department of Internal Medicine, Cleveland Clinic, Cleveland, Ohio
- Aris Persidis, PhD, President/Co-Founder, Biovista: Chairperson/Co-Founder, X-Pandemia; Board of Directors: Biovista, MBFT Therapeutics, X-Pandemia, Charlottesville, Virginia
- Michal Guindy, M.D. MPA, Head of Imaging, Assuta Medical Centers, Tel-Aviv, Israel
- Mihaela van der Schaar, PhD, John Humphrey Plummer Professor of Machine Learning, Artificial Intelligence and Medicine, University of Cambridge

Abstract

AI is transforming many aspects of society, but adoption has been slow in healthcare. The COVID-19 pandemic presented a wide array of challenges and an urgent need for rapid response, which accelerated the development and adoption of AI in the medical domain. This has led to the creation of new tools and methods as well as elevated the need for partnerships between healthcare providers and AI researchers across academia and industry. It has also accelerated interest within the AI research community in applying AI technologies to tackle the unique challenges posed by the medical domain. As a result, the creation of large novel datasets and new technologies has proceeded at an unprecedented pace. This session will focus on lessons learned from the accelerated development and deployment of AI technologies in healthcare in response to some of the urgent needs and challenges and posed by the pandemic which may form the basis of a “new normal” for AI in the medical domain.

Learning Objectives

The specific learning objectives are:
- Review the lifecycle of AI development and adoption in the medical domain
- Identify how regulatory and ethical challenges to AI projects have been affected by the COVID-19 pandemic, including long-term implications
- Discuss how implementation, usability and adoption of medical AI have been prioritized during the pandemic, particularly given heightened clinician burden
- Examine factors that the medical AI community must confront around fairness, bias, privacy and health equity in leveraging AI to improve care both during the COVID-19 pandemic and beyond

General Panel Description
This panel will bring together experts working in different areas of healthcare who will discuss their point of view on the impact of pandemic on the development of AI technologies.

Panelist Presentation Descriptions:

- **Anil Jain** will discuss how the global pandemic has placed increased demands on the clinical community to simultaneously accelerate knowledge discovery, improve clinical outcomes and leverage existing and novel data sets through advanced technologies such as ML/AI while straddling the need to support overburdened clinicians and researchers, develop and maintain AI-savvy talent, and operate within economic and regulatory constraints. He will discuss the possible roles of large companies, small companies, provider organizations, and life sciences in collaborating towards a shared goal through shared risk. As a former member of the US Health Information Technology Advisory Committee formed by the 21st Century Cures Act, he will discuss the role of policymakers to accelerate innovation and promote meaningful adoption of medical AI as well as the role of adjacent technologies, such as Electronic Health Records, in supporting the medical AI community.

- **Aris Persidis** will provide an overview of AI-driven drug repositioning for the Covid-19 pandemic and beyond. The pandemic revealed immediate treatment needs which galvanized the entire industry. Plasmapheresis was the front runner during the wait for vaccines. Then came the vaccines, which must continue to evolve to keep pace with variants, similar to the annual development cycle for influenza vaccines. At the same time, the lack of effective drugs to treat the active infection and its complications (the latter now formally called PASC) became a glaring gap. Drug repositioning emerged as the starting point of choice, as the most efficient way to identify or develop a new drug by leveraging existing ones. AI became a cornerstone of next-gen efforts to reposition existing compounds, instead of starting from scratch, and is becoming a cornerstone in efforts for drug development for PASC and preparedness for future pandemics.

- **Michal Guindy** will review the exponential growth in the number of projects to develop AI in Medical Imaging (MI) for COVID-19 as reflected by the extraordinary number of papers shared in 2020 on the topic. She will discuss the shift in openness to collaboration and the increased number of interdisciplinary and often multi-national collaborations of medical professionals with computer vision researchers. She will present results of an analysis regarding the focus of AI in MI for COVID-19 compared to traditional MI and what characterizes mature studies. Finally, she will present an analysis of patients’ behavior shift during the pandemic and its implication for health systems.

- **Mihaela van der Schaar** will discuss her interactions with governments regarding adoption of proven machine learning methods and use of existing data to help healthcare systems respond to the pandemic, conduct research and statistical analysis regarding the nature of the disease and its spread, and explore the potential impact of machine learning on clinical trials. She will also discuss how evidence is generated regarding the variability of the impact of COVID-19 across ethnicities, which is particularly pronounced within certain countries. Finally she will share results of a study of ethnicity as a COVID-19 risk factor in England.

**Importance of the Topic**

The pace at which AI has transformed the medical world has been slow compared to other domains. The generation of AI technologies often requires long cycles of negotiating data access agreements, harmonizing and cleansing data, developing advanced algorithms and designing and implementing holistic solutions. The sensitivity of healthcare data and associated regulations, as well as the unique infrastructure and incentive systems in the healthcare industry, often further complicate the process. The COVID-19 pandemic presented unique pressure on the AI development process to accelerate delivery of solutions. Vast amounts of healthcare data were quickly generated in care systems as part of the standard monitoring of the increasingly large number of people who have been affected by COVID-19. Healthcare workers on the front lines of the pandemic have experienced burnout, which in turn presented pressure on quick generation of AI based solutions that could help in diagnosis, contact tracing, treatment decisions and monitoring. The novelty of the disease meant that new drugs and vaccines were urgently needed, as well as
accelerated approval of the new COVID-19 indication for existing drugs that proved useful. Time was of the essence, so AI technologies that could accelerate the process of addressing all of these challenges were urgently needed. Moreover, from a public-health perspective, governments have been struggling to find an optimal set of nonpharmaceutical interventions (NPIs) that can help contain or stop the spread of the disease with the least possible harm. Many attempts have been made, and again, rich information has been generated that can be harvested manually or with AI technologies that can also analyze the data to inform approval and policy decisions. Finally, due to the risk of contagion, COVID-19 related confinement and stay at home orders, remote care for both related and unrelated medical conditions, have become a valuable mode of operandum in many settings. This rapid adoption of remote care has presented further urgent need of AI technologies to both evaluate, support and optimize delivery of such services.

The session will benefit from the perspective of almost two years of global experience with the pandemic.

Discussion Questions

While all panel participants agree that advanced AI technologies have the potential to transform healthcare and bring better care at sustainable cost, they have different viewpoints regarding the impact of the pandemic on the acceleration of AI development and adoption.

1. Based on your experience and observations, has the pandemic enabled overcoming regulatory, contractual or other non-technological barriers and accelerated the development of AI technologies? If yes, do you believe these changes are here to stay?
2. Can you share concrete examples of how the pressure of the pandemic has led to investment in AI efforts that were timely and useful in practice? Additionally, do you have examples of how the same pressure may have led to less effective efforts or investments in the development of AI technologies?
3. How has the COVID-19 pandemic affected your own involvement in the development of AI technologies? What are your primary lessons learned from your experiences during the pandemic?

Participation

The panel organizers confirm that all participants have agreed to take part in the panel.
Health Information Exchanges and Public Health Data: Lessons from COVID  
Moderator: Donald W. Rucker, MD  
Panel Participants: Jaime Bland, DNP, RN-BC (Cync Health), Lisa Bari, MBA, MPH (SHIEC), Daniel Chavez, MBA (formerly San Diego Health Connect, HealthTech Solutions) and David Horrocks, MBA, MPH (CRISP)

Abstract

The COVID pandemic has focused attention on what public health data we collect and how we collect and potentially share that data. Historically public health agencies have relied on reporting, typically mandated, of pre-defined data fields specifically tailored to previously known questions about the specific illness (COVID) or to syndromic surveillance of pre-defined symptoms. Today with widespread electronic medical record data and health information exchanges (HIEs) exchanging that data at national scale, we have the opportunity to move from mandated reporting to exchange and re-use of data already captured clinically. This panel discussion features HIE experts who have exchanged public health data at scale throughout the United States.

Panelist Topics

The panel will discuss the role of HIEs in public health data, specifically discussing the learnings from many COVID uses cases. The panel will provide the audience with insights to understand how widespread HIE toolsets can allow us to rethink public health data reporting. Two of the panelists currently manage 5 state HIEs (Jaime Bland and David Horrocks). One of the panelists (Jamie Bland from Cync Health (Nebraska and Iowa’s HIE) will describe how these privacy protecting techniques were used to precisely stratify COVID risk factors. Three of the panelists (Jaime Bland, David Horrocks, Lisa Bari) are working on national efforts to connect HIEs and state-based immunization information systems (IIS). One of the panelists (Dan Chavez) did extensive work connecting the San Diego HIE and CIE (Community Information Exchange) with the San Diego Health and Human Services Agency connecting for Social Determinants of Health. The moderator (Donald Rucker) served as National Coordinator of Health Information Technology and will set the stage by describing Office of the National Coordinator funding of the Star HIE program to better link public health agencies, immunization information systems and HIEs.

Topic Relevance to AMIA Members

The COVID pandemic has highlighted the both the need and the opportunity to rethink public health data collection. Reporting, often mandated, has been the centerpiece of public health data collection for 150 years – arguably starting with John Snow’s identification of the Broad Street water pump has a source of a cholera outbreak. Today multiple local and state public health agencies as well as the Centers for Disease Control and Prevention (CDC) have required reporting for a variety of diseases, lab results, and syndromic symptoms. With the COVID pandemic the Department of Health and Human Services (HHS) including the Center for Medicare and Medicaid Services (CMS) added additional reporting requirements for COVID test results, hospital resource uses such as ICUs and ventilators. These mandated reporting requirements included information on race and ethnicity, which were often missing. Additional reporting systems are currently being introduced to track COVID vaccinations.

This overarching reliance on reporting raises two large questions. The first question is whether this approach provides us the data we need to intelligently understand both the current and likely future pandemics. The second question is whether piecewise unidirectional reporting to public health agencies is the best way to rapidly capture information about and respond to a pandemic in a clinical world where we capture vast amounts of medical data as part of the process of care and have extensive coverage by rich, bi-directional exchange networks already in place.

Answering the first question of what data do we need for a pandemic can start by looking at what is captured today and then considering what is needed. Today state and local health departments collect patient-identified information on specific diseases and lab test results. De-identified versions of these data are sent to the CDC and now to HHS as well. While this information is stored in a variety of databases, none of these databases replicate the
functionality of electronic medical records though that may be an implicit assumption by policy makers. In particular, public health databases obviously can only, at best, store the information that was required to be reported. This means that richer details on illness such as a patient’s comorbidities and social determinants such as race and ethnicity are likely to be missing or incomplete.

Maybe most importantly understanding of a disease requires understanding how that disease can change over time. While de-identified population counts can provide an overall sense of disease trajectory, they cannot answer questions such as what is the range of individual patient trajectories, for example from symptoms to disease treatment and outcome, or from a positive test to a negative test to development of markers of immunity. Transitions in care from home to emergency room to intensive care to nursing home or other pathways that may benefit from specific policy interventions are difficult to track with point-in-time reporting.

The need for individually identifiable information is also present with a vaccination program. It is not enough to collect reports from vaccination centers on how many patients were vaccinated and with what type of vaccine. Critical downstream functions such as nuanced measurement of vaccine complications require a numerator and a denominator. While the FDA’s Vaccine Adverse Event Reporting System (VAERS) captures submitted case reports and will likely capture most obvious complications that are either unique or closely related to the administration of the vaccine, such as system will be unlikely to identify that a vaccine, imaginably, increased asthma events in patients with underlying lung disease or increased abdominal symptoms in patients with inflammatory bowel disease (1). Individual information is also necessary to fully ensure equity in the vaccination distribution process. Such insights require systems that track patients over all of their care, all of their providers, and over time. No state, local or federal public health reporting or data platform can consistently provide such information today.

Successful adoption of informatics solutions requires attention to operational detail. Public health reporting, especially mandated reports, raises many operational issues. The structured data fields captured have to be pre-specified and if additional clinical or other questions arise over time, the reporting has to be redesigned. Almost by definition, mandated reports are after the fact and often they are not integrated into the clinical care process so they require potentially massive amounts of additional resources in order to be implemented (2). When reports are not part of the process of care, data quality is suspect as they do not have the many “eyeballs” looking at clinical data and making sure that the data is accurate. Such reports represent points in time and points of geography so correlation is difficult even if patients are identified and could potentially be followed. As with COVID, the impact of a pandemic can change rapidly – retrospective reporting based on questions built to answer concerns raised by prior clinical outbreaks is inherently unlikely to adequately anticipate differing scenarios arising with future pandemics.

The second question raised is whether piecemeal unidirectional reporting (e.g. electronic case reporting) to public health agencies is the current optimal way to rapidly capture information about and respond to a pandemic in a clinical world where we capture vast amounts of medical data as part of the process of care. Almost all US hospitals and physicians now have electronic medical records. While the CDC syndromic surveillance system does capture selected presentations with a focus on respiratory viruses from over 6,000 sites this data flow still represents a small part of the overall electronic medical data generated in the United States (3). Clinical data is present in electronic medical records but largely absent in the data available to public health agencies.

What are the ways we can enrich the information public health agencies use to protect the public? While there is additional COVID Cares Act funding for the CDC improve and enhance its reporting infrastructure, the use of Health Information Exchanges (HIEs) by many state and local public health agencies offers a partially alternate and a partially complementary way to greatly enrich the real-time availability of broadly-based practice data.

Health Information Exchanges are typically state-wide or large regional organizations designed to allow clinical data sharing among all providers in that geography. Many were started as a response to the Health Information Technology for Economic and Clinical Health (HITECH) Act, enacted as part of the American Recovery and Reinvestment Act of 2009 (ARRA), funding during the Obama Biden Administration. There are approximately 100 HIEs in the United States and they cover approximately 92% of the US population (4). These exchanges include not just doctors and hospitals who have options to join EHR vendor sponsored exchanges or private sector exchanges.
which often focus on serving at-risk (economic risk) Accountable Care Organizations but other providers such as nursing homes, home health agencies and in some cases mental health providers such as group homes and shelters (5).

HIEs have full geographic master patient index functionality so they can aggregate patient data no matter the site of service sending that data. Being able to follow patients and thus diseases over time requires that these challenging to implement and manage patient identification activities be robust. This allows HIEs to serve as the geographic source for truth for demographics need to help with social determinants of health. Today HIEs routinely exchange clinical document summaries, lab results and ADT feeds (admit, discharge and transfer messages) as well as providing analytic tools around population health services and quality measures. By tracking patients over time and geography an HIE can provide insights not available through simple reporting. For example, a reference lab will have a positive COVID test but their staff and software are not well-equipped to accurately capture a patient’s race, ethnicity, comorbidities or medications.

State and local HIEs have a number of valuable attributes that position them to serve as the last-mile for public health data. They are all non-profits with public governance avoiding the likelihood of inadvertently creating another toll-taker disproportionately raising health care costs. Their geography often matches that of public health agencies. They have rich experience in cleaning data, de-duplicating data and normalizing data. Because they cross care boundaries, HIE data has many folks viewing that data which provides ongoing quality control and audit. The panel will also discuss the role of state and local HIEs and their relations to various EHR vendor-based exchange activities.

HIEs are wellpositioned as a source of public health data for state, federal, and local public health agencies. They are also well-positioned to implement clinical data sharing privacy provisions and mitigate the risks of that data being used against patients. The non-profit public governance of most HIEs melds well with the governmental nature of public health agencies and they may serve as public health data utilities in a way similar to other public utilities which are heavily regulated. Modern machine learning data analytic techniques allow privacy protecting transformations known as federated learning, split learning or privacy protected learning. One can imagine use of these software tools with diverse data stores at each of the state and local HIEs providing extraordinarily rich and real-time information so public health agencies have a better prospective ability to detect pandemics however they start.

References


Primary Care Informatics Working Group (PCIWG) Panel:
Demonstrating the Value of a Complex System Problem Solver

Matthew Sakumoto, MD¹, Tiffany I. Leung, MD, MPH, FACP, FAMIA², Ryan Jelinek, DO³, Stephen Morgan, MD, MS, FAAP⁴, Deepti Pandita, MD³

¹University of California, San Francisco, San Francisco, CA; ²Maastricht University, Maastricht, The Netherlands; ³Hennepin Healthcare, Minneapolis, MN; ⁴Pediatrics Associates of Greater Salem, Beverly, MA

What might the attendee be able to do after being in your session?

The intended audience is applied clinical informatics professionals, health system administrators in both academic and community settings, and students who are considering a career in applied clinical informatics.

The attendee will be able to:

- Present and justify the value of the role of a primary care clinical informaticist to their employer
- Identify key variables of importance to the health system and the community it serves
- Quantitatively and qualitatively describe improved communication, outcomes, and impact in their organization

Description of the Problem or Gap

This is a submission under the Academic Informatics theme describing the challenges and remedies to workforce development in the Primary Care Clinical Informatics space.

In 2020, 97% of clinical practices and health systems struggled with negative financial effects due to the COVID-19 pandemic¹. However, the health systems that had prioritized complementary leadership from primary care and clinical informatics prior to the pandemic actually flourished². They had information systems and standards of practice in place for population health management and patient outreach to drive innovations in both COVID-19 response and continuity of non-COVID care. Primary care informatics is valuable for the efficiency and efficacy of healthcare delivery systems and the wellbeing of the community of patients they serve. Primary care informatics enables insight and understanding of clinical workflows from diverse perspectives from developer to end-user. Uniquely, the primary care physician is trained to practice with a whole-person view of the patient within the context of a whole system of care and well-being.

In this panel sponsored by the AMIA Primary Care Informatics Working Group (PCIWG) we aim to demonstrate the specific value of combining primary care clinical skills with applied clinical informatics knowledge and best practices.

Methods

This panel will use a clinical case study to describe key shared principles of primary care and clinical informatics and the impact of a primary care clinical informaticist at the level of the patient and the healthcare system.

Our discussion will center around the New Primary Care Paradigm, a joint statement by leading primary care professional organizations identifying seven shared principles of primary care, where care is Person/Family-centered, Continuous, Comprehensive and Equitable, Team-based and collaborative, Coordinated and Integrated, Accessible, and High-value³. Three of these seven principles of primary care are especially illustrative of the core principles of Primary Care Clinical Informatics: 1) Coordinated and Integrated, 2) Team-based and collaborative, and 3) Comprehensive and Equitable.

Results and Panel Description

Moderator will present the overall problem statement and a summary slide of the three domains to be described across the different speakers' roles.

Moderator: Matthew Sakumoto - Practicing virtual primary care internist and recent graduate of the Clinical Informatics Fellowship. He has primary care and clinical informatics experience in multiple settings including
academic health centers, community hospitals, and industry start-ups.

Panel expertise includes:

1. **Deepti Pandita** - Chief Medical Information Officer (CMIO) and Program Director of the Clinical Informatics Fellowship and practicing internist at Hennepin Health in Minneapolis, Minnesota. She is a C-suite executive clinician with background and experience in Care Delivery, Innovation, Population Health, Data and Analytics and Informatics. Her passion is in using technology to bridge health disparities especially in Medicaid populations in order to transform care and create healthy communities. She will describe how her perspective as an internist in various roles across internal medicine informs her role as the CMIO facilitating Coordinated and Integrated care across her health system.

2. **Tiffany I. Leung** - Assistant Professor at Maastricht University, The Netherlands. As an internist and Clinical Informatics diplomat, she has a background in clinical practice in multiple US academic practice settings, currently in Dutch clinical practice, as well as public and population health, and clinical informatics. Her PhD work (ongoing) focuses on understanding and mitigating job distress among physicians especially when healthcare systems undergo redesign. She will share experiences in international health systems to describe Team-based and Collaborative care in PCI.

3. **Ryan Jelinek** - Current informatics fellow and practicing internist. He will discuss how his experiences with caring for patients in both the ambulatory and acute care settings at a safety net hospital inspire his design and incorporation of comprehensive and equitable care into his clinical informatics initiatives. His current work has focused largely on the equitable provision of telehealth and thinking about how access to care can be improved with better technology and workflow process design.

4. **Stephen Morgan** - Immediate past-chair of the Primary Care Informatics Working Group and practicing pediatrician. He will describe specific career pathways, skill sets, and the value and return on investment a primary care clinical informaticist can bring to an organization of any size, from small practice to large integrated health system.

All participants have agreed to take part on the panel.

The panel will conclude with an open discussion section around the specific questions listed below, as well as audience-submitted questions (by utilizing the conference room microphone if in person, or chat function if virtual):

- As a primary care clinical informaticist what area have you had the largest impact and return on investment (ROI)? (Examples can include financial savings due to increased efficiency and time, improved usability, decreased burnout, and increased satisfaction and retention.)
- How has your primary care background improved provider and patient engagement and satisfaction?
- What is your definition of team-based care and what is different about managing team dynamics in the clinical space compared to the clinical informatics space?
- What primary care principles do you bring to designing and implementing shared savings programs or value based care models?

**Discussion of Results**

The panel will share individual experiences and formulate best practices in describing the business value of the primary care clinical informaticist. To reflect the goals of the Academic Informatics theme, we will also share training experiences from the Program Director and Learner perspectives on data-driven educational programs. We plan to facilitate audience discussion around identification of any missing insights, and how AMIA-PCIWG can utilize this information to further support physician informaticians in their role. Facilitated discussion with the audience may also include developing a survey and repository of sample “job descriptions” and business cases detailing Primary Care Clinical Informatics impact and employment opportunities.

**Conclusion**

We provide a panel perspective on how to support Primary Care physician informaticists in highlighting their unique role and values in healthcare organizations, start-ups and industry. We hope to utilize these insights to encourage others to share their experiences and gather a collection of value statements, descriptors, and successes.

**Attendee’s Take-away Tool**

The attendees should have a good concept of the unique value of a primary care clinical informaticist and specific examples of organizational impact. Through the panelists’ experiences we will demonstrate:
Coordinated and Integrated care will lead to new and innovative primary care delivery models
Team-based and collaborative care can improve provider and patient satisfaction
Comprehensive and Equitable care can improve access and patient outcomes, and ultimately add value to the health system

References


Abstract

Primary Care Informatics (PCI) is essential to the efficacy of healthcare delivery systems and the wellbeing of the patients’ community. Uniquely, the primary care physician is trained to practice with a whole-person patient view in the context of systems-based care and these principles translate into valuable PCI insights.

The speakers will present a clinical case study to describe three shared principles of primary care and clinical informatics and the impact of PCI at the level of the patient and the healthcare system; including themes of clinical care that is 1) Coordinated and Integrated, 2) Team-based and collaborative, and 3) Comprehensive and Equitable.

The objectives of this panel are to: 1) Present and justify the value of the role of a primary care clinical informaticist to their employer 2) Identify key variables of importance to the health system and the community it serves, and 3) Quantitatively and qualitatively describe improved communication, outcomes, and impact in their organization.
Accelerating the Use of Standardized APIs to Support Real-World Data and Real-World Evidence for Health Care and Research

Anita Samarth1, Kevin Chaney, MGS2, Gideon (Scott) Gordon, PhD3, Vera Mucaj, PhD4, Andrew Kress5
1Clinovations GovHealth, Washington, DC; 2Office of the National Coordinator for Health Information Technology, Washington, DC; 3Food and Drug Administration, Washington, DC; 4Datavant, San Francisco, CA; 5HealthVerity, Philadelphia, PA

Abstract

The 21st Century Cures Act (Cures Act), enacted in 2016, promotes and funds efforts in research, precision medicine, drug and medical device development, and addresses issues related to the opioid crisis and mental health service delivery in the United States. The Cures Act also supports adoption and use of real-world data (RWD) to support drug development processes and healthcare decisions by developing guidelines for use of real-world evidence (RWE) to make regulatory decisions, post-market requirements, and new indications for approved drugs. Additionally, the Office of the National Coordinator for Health IT (ONC) Cures Act Final Rule promotes adoption and use of standardized application programming interfaces (APIs) such as the Health Level Seven (HL7®) Fast Health Interoperability Resource (FHIR®) APIs for data exchange between health IT systems and health care applications (apps). This emerging ecosystem of APIs and apps allows developers and data aggregators to expand their product offerings to support life sciences and research communities as RWD and RWE become more prolific. Panelists will bring government, standards, research, and industry perspectives regarding successes, barriers, and challenges of using standardized APIs for the collection and exchange of RWD and RWE to support clinical research, precision medicine, drug development, and other uses.

Introduction

In recent years, there has been an emergence of real-world data (RWD) use to support drug development processes, research, and clinical decision making, enabled and accelerated by increased access to, and sharing of, health data for clinical and research purposes. These trends were further accelerated in 2016, with the 21st Century Cures Act (Cures Act),1 which promoted the acceleration of research into preventing and curing serious illnesses through precision medicine initiatives; accelerates drug and medical device development; attempts to address the opioid crisis, and tries to improve mental health service delivery. The law also contained a number of provisions that promoted greater interoperability and adoption of electronic health records (EHRs) and support for human services programs.

Generally, RWD is observational data related to patient health status and/or the delivery of health care routinely collected from a variety of sources outside regulated clinical trials. Data sources may include EHRs, medical claims and billing activities, product and disease registries, patient-generated data (including in home-use settings) and data gathered from sources that can inform on health status, such as mobile devices.2

The Cures Act additionally provides guidance on the use of real-world evidence (RWE), which states it is “the clinical evidence regarding the usage and potential benefits/risks of a medical product derived from the analysis of RWD.” RWE represents the analytical data that details the use, potential benefits, or risks of a drug obtained from non-traditional clinical trials such as randomized controlled trials and observational studies (prospective and/or retrospective).3 With the use of RWE, the Food and Drug Administration (FDA) can obtain insights on how to best design and operate clinical trials/studies in healthcare settings to resolve unanswered questions and improve treatments.

Concurrently, the Office of the National Coordinator for Health Information Technology (ONC) Cures Act Final Rule (Cures Act Final Rule),4 released in March 2020, seeks to accelerate a shift towards greater access and data exchange by enabling consumers to store, aggregate, use, and share electronic health information (EHI) using standardized application programming interfaces (APIs) and apps of their choice by establishing access to EHI “without special effort” and the ability to exchange EHI with third-parties, including researchers. The Cures Act Final Rule established

128
the Health Level Seven (HL7®) Fast Health Interoperability Resource (FHIR®) standards to achieve this interoperability and data exchange to meet these goals.

While it is possible to extract RWD from EHR data using standardized APIs, gathering RWD from most other data sources is not possible using current standards. In addition, EHR data are insufficient to answer the questions asked by researchers conducting complex analyses trying to leverage clinical data with other RWD to develop RWE in comparison to other research common data models such as: FDA Sentinel Initiative,6 Informatics for Integrating Biology and the Bedside (i2b2),7 Observational Medical Outcomes Partnership (OMOP)8, and Patient-Centered Outcomes Research Network (PCORnet).9 Approaches to gathering sufficient RWD through the use of APIs will continue to emerge as the standards evolve and industry is incentivized to use these standards-based technologies as to bridge the digital divide between health IT systems.

Panel Objectives and Presenters

The aim of this panel is to provide insights from both government and industry perspectives on the use of APIs and third-party apps for the purposes of collecting and sharing RWD and RWE to support clinical decision making and research. The panelists will share findings from federally funded efforts to investigate facilitators and barriers to using APIs to advance RWD and RWE. Panelists representing data integrators will discuss their experiences supporting government, provider, pharmaceutical, and life sciences organizations in the use of real-world data and real-world evidence. The panelists will describe both current examples and future use cases using RWD and RWE to promote a robust research ecosystem that integrates data from a variety of sources, including electronic health records (EHRs).

Mr. Kevin Chaney, MGS (moderator), a Senior Program Manager in ONC’s Scientific Advancement Branch, will provide an overview of ONC’s priorities and their relation to the promotion and use of standardized APIs for scientific discovery, lessons learned and currently funded work in the API and interoperability areas. Mr. Chaney will discuss findings from ONC efforts to review the landscape of API use in the field, specifically to support scientific discovery. He will discuss ONC efforts to develop tools and resources to support improved transparency for providers, patients, researchers, and health IT developers and integrators.

Ms. Anita Samarth (panelist), the CEO of Clinovations GovHealth, will introduce and moderate the session. Ms. Samarth has implemented health IT at over 200 provider organizations and will highlight findings from provider organizations, researchers, and app developers implementing and using APIs to access data contained within EHRs.

Gideon (Scott) Gordon, PhD (panelist), a Senior Health Informatics Officer at FDA for the Office of Strategic Programs in the Center for Drug Evaluation and Research, will discuss FDA activities to standardize data for clinical research, submissions to FDA, and post-market surveillance. A significant aspect of Dr. Gordon’s work includes a focus on RWD derived from health IT and other non-traditional sources for use as an adjunct to data from traditional clinical trials and current pharmacovigilance methods.

Vera Mucaj, PhD (panelist), the Head of Trials and Chief Scientist at Datavant, will discuss their mission to connect health care data from a variety of sources to be used for academic research, pharmaceutical research, and clinical care. Dr. Mucaj’s current role includes a focus on clinical trials and using real-world data connected to trial data to generate real-world evidence in a privacy preserving manner, and to make it useful for retrospective cohort analysis, post-marketing studies, hybrid studies, and complex data analytics.

Mr. Andrew Kress (panelist), CEO of HealthVerity, will discuss current activities to provide a seamless, high-governance, privacy-compliant way to connect and exchange complete real-world data about people across the healthcare and consumer data ecosystem, enabling next-gen analytics and applications. He brings perspectives from HealthVerity’s platform that provides linked, de-identified data for real-world evidence, pharmaceutical companies, payers, clinical trial optimization, and population health.

Panel Discussion Questions

• Describe your experience using real-world data or real-world evidence to support clinical decision making and research efforts. How are you seeing novel approaches to using this data to advance health care and research?
• What do you see as the biggest incentives or barriers to using standardized APIs for the collection and sharing of real-world data and real-world evidence? What types of challenges or barriers are the most difficult to overcome?
• What technical, policy, and/or financial needs or gaps have been identified through your work that limit the adoption of standards-based APIs for the collection and use of real-world data and real-world evidence?
• What educational tools and resources are available or needed to support technology companies as they continue to develop products to support the advancement of real-world data and API use for clinical care and research?

Panel Learning Objectives

• Participants will learn the specific technology, policy, and privacy considerations and industry drivers for the promotion of standardized APIs for the use of RWD and RWE, for promoting data access and sharing for research.
• Participants will have an improved understanding of the emerging use cases for using standardized APIs for the collection and aggregation of RWD and RWE for clinical research and health care.
• Participants will engage with panelists in the discussion and share feedback and experiences advancing the collection of RWD and RWE to support traditional use cases such as clinical trials, as well as research studies and clinical decision making.

Conclusion

Some forms of RWD exists in small data sets, which provides for more rapid, yet less robust research. These data sets do not require extensive data curation and mapping. Conversely, there are larger data sets, such as data collected during clinical trials, that are robust but time-consuming to curate and obtain. Often, such data sets cannot be merged in ways that preserve patient privacy preferences. Some data aggregators have developed the infrastructure and open, scalable technologies to host the data and link to any additional RWD through a robust governance structure that controls access to data sets. In the future, the use of standardized APIs such as FHIR® could improve the ability to collect and disseminate RWD for research purposes, and make such data accessible to research organizations that may not have the resources to acquire it from data aggregators and integrators. The rapidly evolving and expanding RWD and RWE market is blurring the lines between health IT and data products that serve traditional grant-funded academic research and industry-led activities (e.g., clinical trials, pharmaceutical and medical device funded research). While expanded use of multiple data sets found in EHRs used in conjunction with other data for large research studies can be linked and matched to create robust RWD data sets, acceptance of this emerging practice requires understanding and addressing resistance and barriers to use of RWD sources.

Statement of Participation

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

References

2. CDISC Real World Data [Internet]. Austin, TX: Clinical Data Interchange Standards Consortium; 2020 [cited 2021 March 3]. Available from: https://www.cdisc.org/standards/real-world-data
Advancing Interoperability: Analysis of EHR Interoperability Measurement Capabilities

Anita Samarth¹, Amit Trivedi, MS², Steven Lane MD, MPH, FAAFP³, Vaishali Patel, MPH, PhD⁴
¹Clinovations Health + Government, Washington, DC; ²Healthcare Information and Management Systems Society, Chicago, IL; ³Sutter Health, Sacramento, CA; ⁴Office of the National Coordinator for Health Information Technology, Washington, DC

Abstract

Increasing interoperability and use of exchanged data is a national priority established by the Medicare Access and CHIP Reauthorization Act (MACRA) of 2015. Currently, progress is measured primarily through national surveys of hospitals and office-based physicians; however, these data sources do not contain a level of granularity to compare interoperability trends across health information technology (IT) developers and providers using their products. In 2020, the Office of the National Coordinator for Health Information Technology (ONC) launched an initiative to understand how healthcare stakeholders measure the ability to exchange and use external information to improve health outcomes and delivery. This initiative’s findings will provide insights on leveraging electronic health record (EHR) data to measure interoperability. The panel will present findings, including leading practices and types of data used by EHR developers and their customers to measure interoperability. The panel will discuss conclusions about known levels of interoperable information exchange that can be drawn from available measures and reporting capabilities. An important objective of the panel is to engage the informatics community to spur discussion on potential measures of value and utility that provide understanding of progress to advance interoperability.

Introduction

MACRA established a national priority to drive widespread interoperability and to report exchange and use of clinical information to facilitate coordinated care and improve patient outcomes.¹ Subsequently, the 21st Century Cures Act of 2016 (Cures Act) requires the Department of Health and Human Services (HHS) and Office of the National Coordinator for Health Information Technology (ONC) to improve the interoperability of health information.² To date, primary data sources for measurement of interoperable data exchange are limited to national surveys of hospital and office-based physicians.³⁶ While these sources offer general insights on exchange and use of health information, ONC seeks national measures feasible to collect from health IT products that send, receive, query, incorporate, and use information from external products and organizations. In addition, while measures derived from health IT may reduce reporting burden on providers, a deeper understanding of interoperability measurement is necessary prior to drawing conclusions about national levels of interoperable data exchange.⁷⁸ For example, clinical quality measures are relatively mature and are developed, tested, and vetted through national consensus processes and EHR generated measures exist for assessing provider burden⁹; however, detailed interoperability measure specifications are nascent and not available in a format that allows consistent and comparable reporting across different health IT products.¹⁰ Further, making conclusions about the value and utility of exchange is challenging, due to the inability of measures that track volume to also convey context.⁸ Together, these factors constrain the ability of clinicians, health care delivery administrators, researchers, and policy makers to understand the level and impact of health information exchange and support their business and policy decisions.

Understanding Interoperability through Interoperability Measurement

ONC initiated a multi-year endeavor to advance interoperability measurement by developing an in-depth understanding of measures created for health IT users and developers. The project involved an analysis of the current state landscape of interoperability measurement capabilities within EHRs and offers recommendations on strategies and specific tactics that consider real-world experience and value to providers. The project approach consisted of an environmental scan of current state of interoperability measurement, including existing methods used by EHR developers, and an iterative interview approach with various key stakeholders, including health IT users and developers.

The panel will present: 1) an overview of the factors that drove national efforts to understand interoperability measures derived from data within and exchanged with health IT products; 2) findings and insights learned from health IT users and developers about the aspects of interoperability that are currently tracked or measured; 3) lessons learned to implement for health IT developers and users both to support better interoperability, and 4) an overview of possible measurement strategies and insights to support future interoperability tracking and measurement.
Panel Objectives and Presenters

The aims of this panel are: (1) to share findings from ONC’s supported project to identify EHR generated measures of interoperability and (2) to invite participants to provide feedback that can advance the national approaches to measuring interoperability. To engage participants, the presenters will frame the discussion with prepared questions for open discussion amongst health IT users, developers, and other interoperability measurement stakeholders in the audience.

Dr. Vaishali Patel (moderator) leads ONC’s Data Analysis Branch. She is responsible for measuring and reporting on national progress related to interoperability and use of health IT. Through her experiences at ONC and Weill Cornell Medical College, where she evaluated NY State’s health IT and health care delivery transformation initiatives, she will discuss ONC’s role and experiences in obtaining measurement data in support of advancing health IT and interoperability in health care.

Anita Samarth (panelist) is CEO of Clinovations Government + Health with more than 25 years of implementing health IT in the health system, hospital, and ambulatory practice settings at over 200 provider organizations. She serves as the technical project director on a multi-year project for ONC investigating interoperability measurement and will discuss project findings and ongoing efforts to identify and define interoperability measures of value.

Amit Trivedi (panelist) is the Director of Informatics & Health IT Standards at HIMSS with more than 20 years of experience in the certification of health IT products and the development of test criteria and methodology, specializing in interoperability and technical testing of health IT products. He will speak to the elements of health IT standards and interoperability strategy and measurement and discuss insights from EHR Association (EHRA), HIMSS Davies awards winners, and other HIMSS HIE and Interoperability Committee and Community efforts.

Dr. Steven Lane (panelist) is the Clinical Informatics Director for Privacy, Information Security & Interoperability at Sutter Health. As a practicing primary care physician and informaticist, he has been a leader in eliminating barriers to information sharing, optimizing the utility of data exchanged, and leveraging health IT tools available to track numerous interoperability aspects related to his practice. Dr. Lane will discuss his experiences using health IT reporting solutions, and present considerations for uses and limitations of measures derived from health IT data to inform clinical and business decisions.

Panel Discussion Questions

The moderator will address the following types of questions to the panelists and encourage the audience to share their input and experience.

1. What EHR capabilities are available to measure interoperability? How do capabilities differ across different EHR products and reporting solutions?
2. What measures are appropriate to measure at the health IT developer level vs. customer level? Describe the level of insights and conclusions that can be derived from developer level vs. customer level measures.
3. What are the most common transactions, standards, and data that are exchanged between different EHR products? What factors affect whether a particular method is used (e.g., Direct vs. national networks like Carequality and Commonwell)?
4. Where can interoperability investments continue to improve health IT user and developer measurement of exchange? How can tracking and reporting be designed to reduce burden on end-users while still providing accurate information?
5. How should the project’s findings about data tracked by health IT developers and users support ONC’s interoperability measurement strategy?

Panel Learning Objectives

1. Participants will learn about the common interoperability measures that health IT developers use or provide for their customers to assess exchange activity, and the factors that support or present challenges to measurement.
2. Participants will understand how interoperability measure structure and definition can support measurement of inter-and intra-organizational data exchange and measure the types of data that are exchanged.
3. Participants will share perspectives on health IT developer and customer capabilities and challenges related to interoperability measurement in real-world settings and learn how individual customizations and configurations present challenges to interoperability measure comparability.
4. Participants will critique and discuss the advancing interoperability measurement ecosystem, including interoperability measures in use by health IT users and developers.
5. Participants will learn how an understanding, and a measure of how organizations use measurement and use external data can provide valuable data for clinical decision support and care coordination.

Conclusion

The first year of this multi-year initiative supports the goal of understanding the current state of interoperability measurement by health IT developers and users. Health IT national initiatives and federal efforts will be able to leverage the data collected on current interoperability measurement methods and investments to leverage current measurement efforts undertaken by health IT developers alongside national surveys to provide meaningful information for clinicians, health care administrators, policy makers, researchers, and other stakeholders. By identifying the practices and data used by developers to measure interoperability, this project will yield deep insight into the state of interoperability that is currently not gleaned through surveys and provide stakeholders with an understanding of the interoperability capabilities of health IT systems as they work in tandem to improve health outcomes.

Statement of Participation

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

References

What will the attendee be able to do after being in your session?
1. Summarize the latest advancements in AI and machine learning models in terms of ensuring safety, reliability, sensitivity, and precision
2. Understand the components that go into securing provider adoption and ensuring utilization of AI tools in the care delivery space
3. Identify and illustrate the limitations of AI/machine learning for health delivery from a governance and regulatory perspective, and create an evaluation rubric built from best practices of evaluating these types of tools, in order to ensure accuracy, actionability, and usability

Description of the Problem or Gap
The increasing focus of AI on the frontier of clinical applications has led to the accumulation of evidence for clinical decision support across a wide range of specialties such as radiology, oncology, ophthalmology, and critical care. FDA has published action plans for regulating AI/ML-Based Software as a Medical Device (SaMD) and approved more than 160 medical AI products. Meanwhile, due to the sensitivity of clinical AI in producing unintended consequences such as incorrect diagnoses, unnecessary treatments, and racial disparities, there is also a growing concern on how to produce safe, reliable, and trustworthy AI for adaptive clinical decision support (adaptive CDS) systems.

Panel Topics
The overarching goal of this panel is to bring together the thought leaders in the field who have experience in both theory and practice to pinpoint the challenges in creating precise, sensitive, and usable AI tools in the care delivery space, drawing on their experience both in the lab and in practice.

The panel will be entirely discussion-based. The panelist will draw from their own real-world case studies to share lessons learned, including recent evaluations of tools that are implemented in health systems right now, creating best practices for model creation, deployment, provider use, evaluation and ongoing governance and regulation. The invited moderator and discussants are all leading experts who will collectively discuss the pros and cons leveraging studies and live-experience building, deploying, and evaluating CDS tools.

Questions to be answered
- With health data being messy, what strategies does one need to use to measure and tackle bias, ensure reliability and safety, and deliver high precision and sensitivity?
- Once the models are created, how do the insights need to be packaged and delivered to providers to ensure adoption and continued engagement/use?
- After deployment, how should one think about evaluation? Without regulatory bodies requiring certain practices, the evaluation process can look different for each tool, so what are the necessary evaluation steps (rubric) that should be followed?
- What are the safety hazards facing these types of tools? How should regulation and governance of these tools look, and what are the best practices for monitoring and governance?
- Do we have any consensus with regards to the important questions to be solved next and the important steps to take to get to the next level of understanding?
Who Should Attend?
The panel is crafted to provide insights for healthcare professionals and informaticists from all areas of practice interested in real-world evidence for real-time insights in healthcare delivery. Potential attendees include: CMOs, CQO, CMIOs, CIOs, clinical informatics fellows, clinicians (superusers), researchers/faculty, administrators, vendors, health IT professionals, and other AMIA participants.

Organizer and Speaker Qualifications
The organizers and discussants of this panel are healthcare system leaders and experts within care delivery transformation, implementation science, patient safety, informatics, and machine learning/AI. The panel proposal is also based on the gaps identified from a series of events previously organized in AMIA CIC Conference 2021 and AMIA annual symposium 2015-2020 and with the overlapping targeted participants in Applied Data Science for Healthcare Workshop from Association of Computing Machinery (ACM) SIGKDD International Conference of Knowledge Discovery and Data Mining (KDD 2018-2020).

Dr. Eric Topol is the Gary & Mary West Endowed Chair of Innovative Medicine, Scripps Research and the Director and Founder of Scripps Research Translational Institute. His research is on individualized medicine, using the digital technologies to understand each person at the biologic, physiologic granular level to determine appropriate therapies and prevention. An example is the use of pharmacogenomics and his research on clopidogrel (Plavix). By determining the reasons for why such a large proportion of people do not respond to this medication, we can use alternative treatment strategies to prevent blood clots.

Dr. Suchi Saria, PhD is the John C. Malone Associate Professor of computer science, statistics, and health policy, the Director of the Machine Learning and Healthcare Lab, and the founding Research Director of the Malone Center for Engineering in Healthcare at Johns Hopkins. She is also the founder of Bayesian Health. Her research has pioneered the development of next generation diagnostic and treatment planning tools that use statistical learning methods to individualize care. Her work has received recognition in numerous forms including Rambus Fellow, NSF Computing Innovation Fellow, IEEE Intelligent Systems to Artificial Intelligence’s “10 to Watch”, the DARPA Young Faculty Awardee, MIT Technology Review’s ‘35 Innovators under 35’, National Academy of Engineering’s Frontiers of Engineering, the Sloan Research Fellow, National Academy of Medicine’s Emerging Leaders in Health and Medicine, the World Economic Forum Young Global Leader, and the National Science Foundation’s Top Four Contribution to Congress. She is on the editorial board of JMLR and the board of several companies using ML for healthcare.

Dr. Ziad Obermeyer, MD, PhD is the Blue Cross of California Distinguished Associate Professor of Health Policy and Management at the UC Berkeley School of Public Health. He is a physician and researcher who works at the intersection of machine learning, medicine, and health policy. He previously was an Assistant Professor at Harvard Medical School, where he received the NIH Early Independence Award for exceptional junior scientists. He continues to practice emergency medicine in underserved parts of the US. Prior to his career in medicine, he worked as a consultant to pharmaceutical and global health clients at McKinsey & Co. in New Jersey, Geneva, and Tokyo.

Dr. Karandeep Singh, MD, MMSc is Assistant Professor in the Departments of Learning Health Sciences and Internal Medicine at University of Michigan. He is a physician, researcher, and educator interested in studying learning health systems, making new discoveries about disease, and improving patient care through technology. His research areas include natural language processing of clinical notes, risk prediction using high-dimensional data, and mobile health. He is also an assistant professor of Medicine in the Division of Nephrology at the University of Michigan Medical School and has previously served as chief resident, and a nephrology fellowship at Brigham and Women’s Hospital in Boston. He is board certified in internal medicine, nephrology, and clinical informatics.

Dr. Marzyeh Ghassemi is an Assistant Professor at the University of Toronto in Computer Science and Medicine, and a Vector Institute faculty member holding a Canadian CIFAR AI Chair and Canada Research Chair. She is the founder of the Machine Learning for Health group which targets "Healthy ML," with a focus on creating applying machine learning to understand and improve health. She will join MIT's IMES/EECS in July 2021. Professor Ghassemi has a well-established academic track record across
Dr. Pei-Yun Sabrina Hsueh is a health AI leader who has spent more than 10 years in industry building CDS tools between her time at IBM Watson, Viome, and Bayesian Health. Dr. Hsueh is the Past-Chair of the AMIA Consumer and Pervasive Health Informatics Work Group, current co-chair of AMIA Women in AMIA Leadership and Awards Subcommitte, and a Board Member on ACM Practitioners Board. She previously served as the Chair of IBM Research Health Informatics Professional Community. She has authored 20 patents and 50 scientific publications and books in AI/ML for Medicine and Personal Health Informatics and edited for various journals including JAMIA OPEN.

Statement of Participation
Each invited presenter/organizer has confirmed via email that they agree to take part in the proposed panel, and the views expressed in this panel do not represent their institutions and the financial interest.

References

Re-imagining Academic-Private Sector Collaboration to Enhance Digital Health Equity

Urmimala Sarkar, MD, MPH1, Ashwin Patel, MD, PhD2, Everett Crosland, MSC, JD3, David W. Bates, MD, MSc4,5,6, Courtney R. Lyles, PhD1
1University of California, San Francisco, San Francisco, CA; 2InquisitHealth, River Edge, NJ; 3AppliedVR, Los Angeles, CA; 4Brigham and Women’s Hospital, Boston, MA; 5Harvard Medical School, Boston, MA, 6Mass General Brigham, Somerville, MA

Abstract

This panel will provide academic researchers, health informatics professionals, vendor representatives, and clinicians with pragmatic recommendations to develop and sustain cross-sector partnerships to improve digital health equity. The panel includes a primary care physician and health services researcher based in a safety-net health care setting, two digital health entrepreneurs, and a moderator with expertise at the intersection of academia and health informatics. The panelists will describe two research projects that investigate and adapt the fit of health technology for historically marginalized and minoritized populations, including why and how they initiated the projects, what they learned, and how their success may be replicated elsewhere. After participating, attendees should be better able to understand current gaps in consumer health informatics for vulnerable populations, recognize how cross-sector partnerships may be able to address some of these gaps, and formulate recommendations for developing such partnerships in pursuit of equitable digital health innovation.

Introduction/Overview

This panel will provide an overview of digital health disparities and offer pragmatic approaches to addressing them with cross-sector partnerships between academia and the private sector.

Health technology developed in the private sector is critically important to deliver health care, especially during the COVID-19 pandemic.1-3 However, despite high interest,4,5 marginalized and minoritized patients are least likely to access digital tools6,7 due to barriers related to broadband, wifi, and device access as well as usability and accessibility (e.g. literacy, inclusivity, language).8,9 One promising approach to this gap between patient needs/interest and technology is to re-imagine public-private partnerships. Previous research describes challenges of university-industry collaboration and misalignment across stakeholder groups,10,11 making it important to share learnings from successful experiences.

Two academic researchers with expertise in health communications, implementation science, digital, and health equity (Dr. Lyles and Dr. Sarkar) launched an academic incubator, UCSF S.O.L.V.E. Health Tech, in May 2019.12 They invited applications from digital health companies interested in adapting or evaluating their product with a focus on health equity, and selected two companies to partner in a 12-18 month incubation cycle. Dr. Patel and Mr. Crosland represent the two companies that partnered with S.O.L.V.E. Health Tech during this inaugural cycle. As a result of this work, the team generated two peer-reviewed publications,13,14 garnered additional venture capital funding, and initiated two follow-up projects. During this panel, Dr. Lyles and Dr Sarkar will provide insight into the collaboration from the perspective of academics and Dr. Patel and Mr. Crosland will share their perspectives as digital health leaders. Dr. Bates will moderate.

We expect academic researchers, health information technology professionals, vendor representatives, and clinicians will benefit from this panel.

Panel Presentations

1. Courtney Lyles, PhD, Associate Professor of Medicine, University of California, San Francisco, “Background on gaps and opportunities for existing digital approaches to improve health equity” (10 min + 5 min for Q&A = 15 min) Dr. Lyles will provide an overview of existing digital approaches in
healthcare. She will describe how they fall short in meeting needs for patients – particularly those from historically excluded groups such as racial/ethnic minorities, older adults, and those with limited English proficiency and limited health literacy.

2. Urmimala Sarkar, MD, MPH, Professor of Medicine, University of California, San Francisco, “Re-imagining private-public partnerships with an academic incubator” (10 min + 10 min for Q&A = 20 min) Dr. Sarkar will describe the formation of an academic incubator designed to address the gaps that Dr. Lyles introduced. She will provide details on how she and Dr. Lyles created the incubator, what they expected, challenges they encountered (including those related to legal issues such as intellectual property), how they adapted in response to the COVID-19 pandemic, and discuss results.

3. Everett Crosland, MSC, JD, Chief Commercial Officer, AppliedVR, “Academic partnerships for digital health approaches” (5 min + 5 min for Q&A = 10 min) Mr. Crosland is Chief Commercial Officer of a company offering therapeutic virtual reality for patients with serious health conditions in the hospital, ambulatory, and home settings. Mr. Crosland will describe the breadth of academic partnerships AppliedVR has engaged in, their rationale for engaging with S.O.L.V.E. Health Tech in light of their prior randomized clinical trials, and the risks/rewards of the partnership.

4. Ashwin Patel, MD, PhD, Co-Founder and CEO, InquisitHealth, “How clinicians in industry can engage with academic partners” (5 min + 5 min for Q&A = 10 min) Dr. Patel is the Chief Executive Officer and co-founder of a remote, technology-enabled peer mentoring company that delivers chronic disease management longitudinally and one-on-one to patients, through telephonic and smartphone outreach in both in English and Spanish. Dr. Patel will delineate learnings and best practices for partnering with academics, including suggestions for advancing this work with resource-limited delivery system partners during the COVID-19 pandemic, from his unique perspective as a clinician-entrepreneur.

5. Moderated discussion with digital health company partners, “Lessons learned from partnering with an academic incubator to improve digital health equity” (moderated by David W. Bates, MD, MSc, Professor of Medicine, Harvard Medical School, Chief of the Division of General Internal Medicine and Primary Care, Brigham and Women’s Hospital) (20 min + 15 min for Q&A = 35 min) Dr. Bates is Medical Director of Clinical and Quality Analysis, Information Systems at Partners Healthcare System, Inc. and a national leader in patient safety, informatics, and digital health equity. He will moderate a discussion between Dr. Patel and Mr. Crosland regarding their experiences working with an academic incubator to adapt and evaluate their digital health products with a focus on health equity. The panelists will provide examples of cross-sector partnerships from the company perspective, challenges they encountered, facilitators, and recommendations for other vendors.

Audience Discussion Questions

1. How can academic-private sector partnerships like these improve health equity? What can other stakeholders (e.g. health systems, clinicians, information technology professionals, and policy makers) do to support this work?

2. What are the biggest barriers and facilitators to establishing partnerships among academic research teams and digital health companies? How do these barriers and facilitators change based on setting and company stage?

3. What are best practices for deciding who to partner with when establishing academic-private sector collaborations?

4. What are examples of pressing topics that academic-private sector partners can study together?

Statement Confirming Participation

As organizer of this panel, I, Urmimala Sarkar, confirm that all panelists have agreed to take part.

References


Broadband Internet Access as a Social Determinant of Health
With Impacts on Health Disparities During COVID-19

Marianne Sharko, MD, MS1, Natalie (Nat) Benda, PhD1, Tiffany Veinot, PhD, MLS2,
Cynthia Sieck, PhD MPH3, Jessica S. Ancker, PhD, MPH4

1. Weill Cornell Medicine, New York, NY; 2. University of Michigan, Ann Arbor, MI; 3. The Ohio State University College of Medicine, Columbus OH; 4. Vanderbilt University Medical Center, Nashville TN

Abstract Disparities in broadband internet access is an important public health issue. Broadband internet access (BIA) affects “the health of people and communities where they live, learn, work and play,” and thereby should qualify as a social determinant health as defined by the American Medical Association. During the COVID-19 pandemic, lack of universal internet access has contributed to the widening gap in health disparities with vulnerable populations disproportionately experiencing the effects of insufficient BIA. In this panel, we will demonstrate how poor internet access influences health disparities, particularly through limited access to healthcare and health-related information. This presentation will highlight the need to identify internet access as a social determinant of health. After participating in this session, attendees should be able to identify the effects of limited BIA access on health disparities and inequities and discuss policies for addressing the disparity.

Keywords: Social determinants of health, broadband internet access, health disparities, telehealth

Introduction The unprecedented COVID-19 pandemic has widened health disparities; this has resulted in as yet unappreciated impacts—particularly on vulnerable populations. As Americans have become reliant on the internet for most aspects of our lives, limitations in access can influence multiple domains of social determinants of health, including economic stability, education, food security, social support and physical environment.1,2 Inconsistent or low-quality internet access also contributes to inequitable healthcare and access to health-related information, which widens existing health disparities. (Figure 1) Vulnerable populations, such as those from low income neighborhoods, older adults, and racial and ethnic minorities, have already experienced deep-seated inequities in health care which have been compounded by gaps in information and healthcare access.3,4 There are trends in disparate infections rates. For example, in Milwaukee County, WI, 81% of the deaths reported from COVID-19 in were Black residents, despite a Black population of 26%.5 Blame for this disparity is often placed on underlying comorbid conditions such as obesity and hypertension; however, there are structural factors that are also contributory. Differences in resources accessible to people who have better broadband internet access enable them to work from home and more effectively access healthcare and information. During the pandemic, those without in-home internet access had a limited ability to gain information about office closures, schedule appointments, and ultimately access care. Those we may have been able to access a healthcare provider via telephone-only missed the ability to have the health professionals’ complete visual examinations, that are possible to some extent via video-based visits. Issues are also not limited to those without an in-home internet connection, but others who had unstable internet service, or had to contend with family members working or schooling from home, also faced barriers to accessing video-visits. Lastly, vital public health information has changed rapidly during the pandemic, and predominantly exists online. COVID-19 vaccine administration has further demonstrated inequities for those without stable broadband internet access, as many states are showing Whites
have been vaccinated at a much higher rate than African Americans who have borne the brunt of COVID-19 related disparities.\(^{(7)}\)

In 2017, the American Medical Informatics Association urged the Federal Communication Commission to consider BIA as a SDOH.\(^{(8)}\) This would identify the need to provide universal access to internet access in order to work towards an equitable healthcare system. In recognition of this issue, the Federal Communications Commission recently approved the $3.2 billion Emergency Broadband Benefit Program during the pandemic. This program provides support for broadband services to help connect low-income households during the COVID-19 pandemic.\(^{(9)}\)

This panel will explore how limitations in internet access contribute to inequalities in access to healthcare and health information. Our panel members will provide an overview of social determinants of health and will use their research to demonstrate how limitations in BIA can affect health disparities by limiting access to healthcare and health information. We will conclude by discussing policy implications for ensuring that public programs include BIA beyond the COVID-19 pandemic.

**Discussion questions**
- What would be the benefits or drawbacks of defining broadband internet access as a social determinant of health?
- How can we improve broadband internet access and where do we focus these improvements for maximum impact?
- How can we identify populations in need of improved broadband internet access?
- What policy mechanisms exist to improve broadband internet access?

**The Panel**

Each panelist will share research that seeks to provide information on the importance of recognizing limited BIA as a social determinant of health.

**Dr. Natalie (Nat) Benda (moderator)** is an expert in human factors engineering/systems safety, and her work involves advancing health equity and inclusivity through technology-based interventions. She will serve as the moderator. Following the panelists’ presentations, the moderator will facilitate discussion by first posing one to two questions to the panel regarding policy or other mechanisms for addressing BIA disparities. The audience will then be invited to ask additional questions and engage in a discussion surrounding the role of informatics in promoting BIA and ensuring it is provided as a public resource.

**Dr. Jessica Ancker** has conducted extensive research on health literacy and numeracy as social determinants of health and on the use of health information technology by diverse patient populations. She is leading a comparative effectiveness study funded by PCORI to examine the impact of the COVID pandemic on telemedicine uptake and impact. She will provide an overview of public health research on the impact of social determinants of health. She will introduce the concept of limited broadband internet access as a social determinant of health and provide a context for the need to ensure equitable access to this resource in order to reduce health disparities. Her overview of broadband internet access issues will be supplemented with evidence from her ongoing PCORI study related to the utilization of telemedicine during the pandemic.

**Dr. Marianne Sharko** is a pediatrician who received her master’s degree in Health Informatics from the Population Health Sciences Department at Weill Cornell Medicine and has completed a Preventive Medicine Residency. She is currently working in a pediatric obesity clinic and conducting research through the Empire Clinical Research Investigator Program (ECRIP). Her research interests include the intersection of obesity and social determinants of health and how this synergy contributes to health disparities. Dr. Sharko is currently conducting an interview study discussing the pandemic experience of families in a low-income population on public health insurance. She will share their words and experiences regarding BIA limitations and how it has affected their lives. She will identify challenges of insufficient BIA and sources of support to improve BIA or to compensate for insufficient access.

**Dr. Cynthia J. Sieck** is an Associate Professor in Family and Community Medicine in The Ohio State University College of Medicine and in the Center for the Advancement of Team Science, Analytics, and Systems Thinking (Catalyst). Her research focuses on patient engagement in health care, particularly through the use of health information technology (HIT), and disparities in health care access and outcomes. With a background in Health Behavior and Health Education, she studies how patients can increase their level of engagement to fully participate in their own health care, and how healthcare systems can best support patients in these efforts. She will discuss efforts to identify populations in need of improved broadband internet access and community collaborations that support digital skill and literacy training to ultimately improve access to health services and health information.
Dr. Tiffany Veinot is Associate Dean for Faculty at the University of Michigan (UM)’s School of Information. She is also Professor at the Schools of Information and Public Health at UM. Her research focuses on “community health informatics,” or the use of information systems and services to improve the health of marginalized populations and reduce health disparities. She will discuss findings from an NSF-funded project focused on telehealth access at a federally qualified health center with six clinical sites in Metropolitan Detroit. Detroit is 78.6% African American, and Detroit households also have lower rates of in-home broadband internet (59.3%) when compared to the US (80.4%). While placing the findings in the context of structural racism and digital redlining, Veinot will discuss analyses of technical problems during video visits with FQHC patients, including high numbers of visits interrupted or restarted—or switched to telephone—due to technical challenges. She will also outline results of a pilot intervention in which trained master’s students helped patients get set up for telehealth, including challenges and problems encountered in doing so. Finally, she will outline ongoing efforts to quantify and reduce the complexity of telehealth processes and systems in order to reduce barriers to access for FQHC patients.

Learning Objectives
After participating in the session, the attendant should be able to:

- Describe how limited BIA contributes to widening health disparities
- Appreciate the need to identify BIA as a social determinant of health
- Advocate for strategies to widen BIA and reduce barriers to access for vulnerable patient populations

Panel Organizer Statement
All participants have agreed to participate in this panel.

Conflicts of Interest
The participants have no conflicts of interest to disclose.

Acknowledgements
Jessica Ancker and Natalie Benda are supported by PCORI COVID-2020C2-10791 (Ancker and Kaushal, co-PIs). Marianne Sharko is supported by the NYS Department of Health Empire Clinical Research Investigator Program. Tiffany Veinot is supported by NSF RAPID-COVID-19-#2031662 and a Google Health Unrestricted Gift.

References
Informatics Career Journeys Among Academic, Government and Industrial Roles

Moderator:
Jane L. Snowdon, PhD

Proposed Panelists:
Patricia Flatley Brennan, RN, PhD
Judy Murphy, RN
Marion Ball, EdD
Yull Arriaga, MD

Proposed Panelist Affiliations:
a IBM Watson Health, Cambridge, MA
b National Library of Medicine, Bethesda, MD
c Freelance, Lake Elmo, MN
d University of Texas at Arlington, Arlington, TX
e Bristol Myers Squibb, Lawrenceville, NJ

Abstract: The healthcare industry is transforming along four dimensions: societal, economic, individuals, and organizations. First, scarce resources and rising demand are driving a relentless focus on cost reduction and quality with improved and equitable access. Second, patients and families are empowered with digital tools such as smart devices, wearables, telemedicine, and remote monitoring that are driving new expectations in standards of care and ownership of health interventions and outcomes. Third, business and operating models are driving consolidation and disruption, new entrants, and an evolving regulatory environment.

Having healthcare informatics experience in various sectors (e.g. academic, industry, and government) can provide a broader perspective about the challenges of and opportunities for developing and applying informatics and innovative technologies to facilitate healthcare transformation and collaboration leading to advances in science. In this panel, we present the experiences of several individuals with successful informatics careers that involved transitions from academia to government, government to industry, industry to academia, and academia to industry. The panelists will share their lessons on how to be successful in a dynamic environment, support for research, growth opportunities, skills development, and how their time in one sector influenced their career choice.

Moderator and Proposed Panelist Information:
- Jane L. Snowdon, PhD, FAMIA is Deputy Chief Science Officer, Scientific Operations, Center for AI, Research and Evaluation, IBM Watson Health
- Patricia Flatley Brennan, RN, PhD, FAIMBE, FACMI, FAAN is Director of the National Library of Medicine
- Judy Murphy, RN, FACMI, FHIIMSS, FAAN is a Nurse Executive and Health IT Leader who served as ONC Deputy National Coordinator for Programs and Policy
What might the attendee be able to do after being in your session?
Attendees at this session will be better able to navigate informatics career transitions between academic, industry, and government sectors. Panelists will share insights into how to be agile, adaptive, and successful in dynamic healthcare sectors and environments.

Description:
The 90 minute panel will be organized as follows:

- Dr. Snowdon will introduce the topic, the panelists, and their organizations (15 minutes)
- The panelists will discuss each of the prepared topic questions below (45 minutes)
- Dr. Snowdon will facilitate questions from the audience (30 minutes)

- What factors influenced your decision to go into the [academia, government, or industry] sector as an initial career choice?
- What were the most important factors you considered when evaluating a career transition to your next opportunity?
- How have you employed informatics and technology to address the challenges and opportunities presented by the three, global disruptive forces above for your sector?
- How have your multi-sector experiences helped to facilitate collaborations leading to advancements in science?
- Which skills do you recommend acquiring to become agile and adaptive, and to keep pace with change in your sector?
- How have your experiences and skills development prepared you and influenced you to make a career transition from one sector to a different sector?

Participation statement:
All proposed panelists are aware of this panel submission and have agreed to participate in the panel if the proposal is accepted.
Home Healthcare to Primary Care Data Interoperability Challenges and Opportunities

Paulina S. Sockolow, DrPH, MS, MBA1, Kathryn H. Bowles, PhD, RN, FAAN, FACMI2,3, Maxim Topaz, PhD, RN, 3,4, Edgar Chou, MD, MBA, FACP5
1Drexel University, Philadelphia, PA; 2University of Pennsylvania School of Nursing, Philadelphia, PA, 3Center for Home Care Policy & Research, Visiting Nurse Service of New York, NY, 4 School of Nursing and Data Science Institute, Columbia University, New York, NY, 5Thomas Jefferson University, Philadelphia, PA

Abstract

Medicare patients are expected to have a primary care office visit within 2 weeks of hospital discharge. Over 1 million of these patients receive home healthcare (HHC) services. Despite HHC electronically collecting a wealth of patient assessment data, primary care receives a fax with little data asynchronous to the visit. Informatics has opportunities to address the interoperability challenge to improve patient safety and outcomes. This panel discusses challenges and opportunities in information continuity for primary care to have the right HHC information at the right time. Panelists will also discuss applying predictive analytics approaches with HHC data to identify rehospitalization risk factors, which if communicated to primary care could trigger preventive interventions. Session learning objectives are: describe visit and HHC information; identify HHC to primary care interoperability barriers; discuss how HHC data can inform primary care of patient risk factors; and describe risk prediction models using HHC data.

Keywords: communication, home health care nursing, home health nursing, primary health care, continuity of patient care/standards, transition of care, nursing informatics, documentation.

Introduction

This panel of informatics experts will discuss the need to improve the interoperability and sharing of patient data from home healthcare (HHC) to primary care. Panelists will describe the relevance of HHC data to the patient-centered primary care team, presenting as exemplars the rich HHC data used to develop valuable prediction models. Panelists will discuss the challenges of unleashing patient data trapped in HHC electronic health record systems (EHR) and delivering it to the primary care team. The intended audience is primary care providers, nurses, EHR vendors, and informatics researchers. The objectives are to: 1. illuminate a gap in information continuity from HHC to primary care; 2. highlight an opportunity for the patient-centered primary care team to have the right information at the right time to improve patient safety and outcomes; and 3. encourage informatics research in HHC and primary care interoperability.

This topic is especially timely and important as the Centers for Medicare and Medicaid (CMS) encourages healthcare organizations to move towards value-based care where hospitalizations and readmissions are key quality and cost drivers. To prevent hospital readmissions and to care for the growing chronically-ill population, the United States increasingly uses HHC. More than 3.4 million homebound Medicare beneficiaries, mainly chronically ill, older adults are admitted to HHC services annually. Of these, one-third transition to HHC from a hospital.[1] Evidence based recommendations to reduce rehospitalization indicate the primary care team should evaluate HHC patients within two weeks of hospital to HHC transition.[2] Outpatient follow-up within one week, in combination with timely HHC nursing visits, has been shown to reduce rehospitalization in high risk heart failure [3] and sepsis [4] populations. Correspondingly, CMS Transitional Care Management requires a patient office visit within 14 days of hospital discharge, the Transition of Care (TOC) visit.[5]

For the TOC visit, the primary care team receives a minimal amount of non-electronic HHC patient data. However, HHC clinicians collect a rich set of structured data about the clinical, functional, and service needs of the patient during the first (admission) home visit. Guided by the CMS required Outcome and Assessment Information Set (OASIS),[6] HHC clinicians assess nearly 100 items including medications, patient safety, wounds, and cognitive and functional capabilities. Unfortunately, only a tiny subset of the structured OASIS data and narrative clinical notes are faxed to the primary care physician in the request to sign-off on HHC orders necessary for reimbursement. The faxed form (known as form ‘485’), contains a paucity of structured data, a preponderance of narrative text, and is not received in synchrony to the patient’s outpatient follow-up visit. Due to information silos and lack of implemented interoperability standards, there is a serious communication deficit along the HHC and primary care transition in care continuum.
Unleashing data from the HHC EHR is critical to capitalize on predictive analytics that may identify hospitalization risk factors, which if communicated to the primary care team could trigger preventive interventions. Also, the brief 15-20-minute TOC visit limits the amount and quality of information primary care providers can gather as they address acute and chronic illnesses, and hospitalization risk factors, making communication from HHC all the more valuable.

Primary care and HHC information siloes are reinforced by a lack of implementation of interoperability standards between the disparate HHC and primary care EHRs. Communication of OASIS data not on the 485 is not fully supported by existing interoperability standards. For example, the U.S. Core Data for Interoperability (USCDI) Version 2 [7] does not include medication self-administration capability. HHC is explicitly excluded from Advancing Care Information (formerly Meaningful Use) and implicitly excluded from USCDI due to health information exchange participation requirements. Information communication discontinuity creates potential missed clinical opportunities (not having information in the right place at the right time) which can impact patient outcomes. Implementation of an interoperability standard that supports communication of specific OASIS data and risk indicators to the primary care team for the TOC visit is necessary to address the fragmented delivery of care across the health care continuum.

**Panel Moderator: What Are the Interoperability Challenges in the Current Health Care System?**

**Paulina S. Sockolow DrPH MS MBA** will open the panel with a short overview of the primary care TOC visit, patient information collected in HHC, and HHC patient information communicated to primary care. She will discuss interoperability between HHC and primary care EHRs, and challenges of the USCDI interoperability standard.

Dr. Sockolow, Associate Professor at Drexel University, just completed an observational study of information availability and decision-making during the HHC admission.

**HHC Data for the Primary Care Team’s Transition of Care Visit: Challenges in Interoperability**

**Edgar Chou MD** will identify HHC data available for the TOC visit and gaps in data availability relevant to patient-centered, value-based care. He will detail the intersection of OASIS data, TOC data, and the USCDI. Dr. Chou will discuss HHC data that would be informative for the TOC visit including OASIS data (e.g., cognitive function, medication self-administration capability) and risk indicators (i.e., hospitalization, falls). He will specify this data’s presence or absence in the USCDI and offer research recommendations on USCDI HHC data inclusion.

Dr. Chou, Clinical Associate Professor at Thomas Jefferson University, is a primary care physician and an informatician. His research interests including leveraging information technology to improve the quality and efficiency of care and health care disparities.

**Structured HHC OASIS data: An Opportunity to Communicate Fall Risk**

**Kathryn H. Bowles PhD RN FAAN FACMI** will provide detail about the rich, structured, and standardized HHC data captured in the OASIS. She will describe the use of OASIS data in a machine learning analysis to predict fall risk, a cause for hospital admission, from her recently completed study.[8]

Dr. Bowles is Professor and van Ameringen Chair in Nursing Excellence at the University of Pennsylvania School of Nursing. She is the Vice President and Director of the Center for Home Care Policy & Research at the Visiting Nurse Service of New York. Dr. Bowles’ programs of research are in transitional care, decision support, HHC, and the EHR.

**Narrative HHC data: An Opportunity to Communicate Hospitalization Risk**

**Maxim Topaz PhD RN** will describe the HHC clinical notes data. He will show how this data can be used in predictive modeling with examples from his NLP studies to identify hospitalization risk factors in clinical notes.[9-11]

Dr. Topaz is an Associate Professor at Columbia University School of Nursing, Columbia Data Science Institute, and Research Scientist, Center for Home Care Policy & Research at the Visiting Nurse Service of New York. He conducts research in health informatics, focusing on natural language processing (NLP) and clinical decision support in HHC.

**Learning Objectives**

After participating in this session, attendees will be able to:

- Describe TOC information and HHC information
- Identify HHC to TOC interoperability barriers
- Discuss how HHC data can inform the primary care team and HHC nurses of patient risk factors
- Describe a couple of informatics approaches to develop risk prediction models with HHC data
Panel Organizer Statement

All participants have agreed to participate in this panel. The AMIA Primary Care Informatics Working Group and the Nursing Informatics Working Group endorse this panel.

Conflict of Interest

The panelists have no conflicts of interest to disclose. The HHC research of Dr.s Topaz, Sockolow, and Bowles is funded by the National Institute of Nursing Research (NINR) (R01NR018831 PI: Topaz). Dr. Bowles is also funded by the National Institute of Aging (R01 AG056607; P30 AG064105), the Agency for Health Care Quality and Research (AHRQ) (R01HS026599), NINR (R01 NR018196), National Institute on Minority Health and Health Disparities (R01 NR019315), and the Betty Irene Moore Foundation. Dr. Topaz and Bowles are also funded by AHRQ (R01HS027742).

Discussion Questions

- Are primary care teams aware or knowledgeable of the OASIS and HHC hospitalization risk factors information? If aware, would they like to know the OASIS information; the hospitalization risk factors information? If not, how to address this deficit?
- What OASIS data elements would primary care teams incorporate in the TCM visit (e.g., Activities of Daily Living, social determinants of health)?
- What changes to inclusion of data elements in the USCDI interoperability standard would support data sharing between HHC and primary care EHRs?
- What changes to HHC and primary care EHRs are needed to adopt the revised interoperability standard?

References

Applying State of the Art Language Models to Enable Better Clinical Natural Language Processing

Bryan D. Steitz, PhD¹; Emily Alsentzer, MS²; Hoo Chang Shin, PhD³; Byron C. Wallace, PhD⁴; Adam Wright, PhD¹

¹Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN
²Massachusetts Institute of Technology, Cambridge, MA
³NVIDIA Corporation
⁴Northeastern University, Boston, MA

Abstract

The natural language processing (NLP) domain has seen important advancements in recent years. Transformer-based contextual word embedding models, such as BERT, have demonstrated significant improvement in NLP performance across a range of tasks. Similarly, the growing scope of data has enabled large language models that support efficient transfer learning. The Bidirectional Encoder Representations for Transformers (BERT) model has demonstrated promise in the clinical domain, but research on the degree of model training and configuration necessary for clinical tasks is in its infancy. This panel will discuss several projects that highlight state-of-the-art training and development of bidirectional language models across clinical and biomedical contexts. The panel will discuss differences between models and vocabularies trained on various specialty texts, including scientific literature and clinical documents. The audience will engage in discussion about opportunities for future model development and privacy concerns when sharing clinical vocabularies and model parameters derived from potentially sensitive data.

Learning Objectives

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

- Describe how contextualized word embedding models can improve NLP performance
- Understand and contrast different “pre-training” strategies, and which might be best for clinical texts
- Weigh the trade-offs and benefits of different pooling strategies to derive aggregate representations from BERT-based vectors induced over chunks of input
- Compare and contrast bidirectional transformer models based on the dataset on which they were developed and trained
- Evaluate practical opportunities for transfer learning

Panel Description

Developing meaningful, rich word representations is integral to creating accurate natural language processing (NLP) models. Word embedding approaches, such as Word2Vec, have shown promise in improving NLP performance since their introduction in the early 2010s. These traditional approaches take a context-independent view of text, which limits their ability to capture semantic meanings. Recent advancements have developed large context-dependent word embedding models using large text corpora, such as Bidirectional Encoder Representations from Transformers (BERT), which have demonstrated significant improvements to NLP tasks. Despite the promise of these models, the degree to which generic architectures and weights can be adapted to specific application domains (such as healthcare) is an active area of research.
Since the inception of bidirectional transformers, developers and researchers have created numerous task- and domain-specific models using two approaches: pre-training and fine-tuning. Pre-training “in-domain” entails training a model on domain-specific text (e.g., clinical notes from EHR) to learn to induce contextualized word embeddings. Domain-specific models have demonstrated improved performance, but the training task comes at significant computational cost. In contrast to pre-training, fine-tuning tasks allow researchers to apply transfer learning using existing pre-trained models. In fine-tuning, one typically adds modules (e.g., a single linear layer) on top of an existing architecture to add domain-specific context to the model with significantly less computational cost and a smaller corpus of data than required for pre-training. Understanding the tradeoff of pre-training and fine-tuning that are necessary for domain-specific tasks is an active area of research. Further, many of the pre-trained models have maintained generalized vocabulary, which was created using a non-domain-specific corpus. The degree to which creating new vocabularies for domain-specific models is currently unknown. These areas of research are rich for novel discovery and represent significant opportunities to improve clinical and biomedical NLP tasks.

For this panel, we selected a set of research projects that feature cutting edge use and development of bidirectional transformer models in the clinical and biomedical contexts. The panelists will detail the challenges that they are solving, their approach advancing the state-of-the-art, and key results. The presentations by each panelist will be followed by a moderator-led interactive discussion around developing task-specific architectures, transfer learning, model dissemination, and future work in model development for clinical contexts. The panelists represent different institutions and have diverse background and training.

**Bryan Steitz, PhD** (panelist and organizer) is a postdoctoral research fellow in the Department of Biomedical Informatics at Vanderbilt University Medical Center. Dr. Steitz will present about his work to apply bidirectional transformers to understand clinical workflow tasks. Dr. Steitz’s research interest focuses on using NLP to understand the context of clinical messages to reduce alerts and improve information seeking in the EHR. His work has discovered and classified the content of messages across clinicians and staff treating breast cancer patients. He identified high degree of triage work for clinicians when communicating about logistical issues, which suggests an opportunity to improve message routing.

**Emily Alsentzer** (panelist) is a PhD student in the Health Science and Technology program at Harvard Medical School and MIT. She received a B.S. in computer science and M.S. in biomedical informatics at Stanford University. Her research involves developing NLP and graph neural network methods for healthcare with applications to summarization and rare disease diagnosis. She will discuss her work to develop clinicalBERT, a publicly available BERT model adapted for the clinical domain and outline promising directions and challenges for leveraging clinical language models.

**Hoo Chang Shin, PhD** (panelist) is a senior research scientist in conversational AI at NVIDIA. Dr. Shin is an expert in applying machine learning and deep learning approaches to contextualize unstructured medical data. His work in NLP has focused on understanding the training and transferability considerations of large language models. As a panelist, he will discuss his work on BioMegatron – one of the largest published language models, which was trained on a corpus of general and domain-specific text. He will also discuss the trade-off between model training time and model size.

**Byron C. Wallace, PhD** (panelist) is an Assistant Professor at Northeastern University. Dr. Wallace is an expert in developing NLP methodologies, with a particular interest in model interpretability, trustworthiness of model outputs, and active learning. As a panelist, Dr. Wallace will present about his recent work to understand the privacy implications of training large language models over clinical notes with an emphasis on how to safely distribute these models. He will also discuss trade-offs and takeaways observed in working with BERT-based models to process biomedical literature.

**Adam Wright, PhD** (moderator) is a Professor of Biomedical Informatics at Vanderbilt University Medical Center and the Director of the Vanderbilt Clinical Informatics Center. Dr. Wright is an expert in clinical informatics and has extensive experience applying language models to process clinical documents.
in support of a variety of clinical applications. As the moderator, Dr. Wright will kick off the panel, introduce panelists, and provide a high-level overview of the BERT architecture. He will frame the discussion associate the panelists’ presentations within the current state of BERT language model research. He will also facilitate participatory discussion between the panelists and audience.

**Importance of this panel**

The growing scope of data, coupled with significant advancements in contextual embedding architectures, yields significant promise to the informatics community for improving NLP performance. The AMIA community is a leader in the application and development of NLP to clinical contexts. It is important to share current results and approaches as well as engage in thoughtful discussion about future directions and opportunities to further improve these models. The anticipated audience for this panel will likely cover a broad spectrum of the AMIA community who are building or applying NLP models to clinical and biomedical contexts, including academic and industry researchers, healthcare leadership, and biomedical informatics trainees.

**Discussion Questions**

1. How much data is necessary to retrain from scratch?
2. To what extent does growing model size affect the need for more specific contextual models?
3. To what extent can models trained on data containing patient identifiers be shared? Is there any research that looks into reconstructing clinical text using these models?
4. How can we evaluate large language models in the clinical context?
5. Are there confidentiality issues introduced by sharing large language models?
6. What are the ethical issues related to the use of large language models? (E.g. environmental impacts and bias).

**Statement of Agreement to Participate**

All participants have agreed to attend the 2021 AMIA Annual Symposium and participate in this panel.

**Acknowledgement**

Work presented by Bryan Steitz and Adam Wright was supported by grant 5R01AG062499 from the National Institute of Aging.

**References**


Career Development Issues for Women in Biomedical Informatics within Professional Organizations

Donghua Tao, PhD, MA, MS¹, Duo (Helen) Wei, PhD², Rubina F. Rizvi, MD, PhD³, Deepti Pandita MD, FACP, FAMIA¹⁴, Bushra Alghamdi MS, PhD candidate⁵, Polina Kukhareva, PhD, MPH⁶, Margarita Sordo, MSc, PhD, FAMIA⁷, William Hersh, MD, FACMI, FAMIA, FACP⁸, Omolola Ogunyemi, PhD, FACMI ⁹,¹⁰, Kelly Taylor, BS¹¹, Gretchen Purcell Jackson, MD, PhD, FACS, FACMI, FAMIA³,¹²

¹Saint Louis University, St. Louis, MO; ²Stockton University, Galloway, NJ; ³IBM Watson Health, Cambridge, MA; ⁴Hennepin Health Care, Minneapolis, MN; ⁵Case Western Reserve University, School of Medicine, Cleveland, OH; ⁶University of Utah, Salt Lake City, UT; ⁷Brigham and Women’s Hospital, Harvard Medical School, Boston, MA; ⁸Oregon Health & Science University, Portland, OR; ⁹Charles R. Drew University of Medicine and Science; ¹⁰University of California, Los Angeles; ¹¹AMIA, Rockville, MD; ¹²Vanderbilt University Medical Center, Nashville, TN

Abstract

In 2020, the Women in AMIA(WIA) Networking, Mentoring, and Lifecycle Committee administered a survey to: 1) identify resources/needs to help support women in informatics and to connect them to the appropriate resources; and 2) further the opportunities and attract more women into the informatics field. The survey results revealed important career development issues for women and roles of professional organizations in nurturing a culture of diversity and inclusivity, providing professional skills training, building networking and connections, and improving communications within and between organizations and members. After this panel, the audience should be able to:

- Understand the career development challenges and issues for women in the biomedical informatics field.
- Exchange ideas and suggest solutions targeting female professionals in biomedical informatics for supporting their career development.

Panel Description

One of the overarching themes in the 2020-2025 American Medical Informatics Association (AMIA) strategic plan states that “AMIA will harness the diversity inherent in the informatics profession and the initiatives we champion, and will bring it to life within AMIA.”¹¹ Aligned with this perspective, in 2020 the WIA Networking, Mentoring, and Lifecycle Subcommittee administered a survey to current AMIA members, both male and female, in different career stages and work settings, to: 1) identify resources/needs to help support women in informatics and to connect them to the appropriate resources; and 2) further the opportunities and attract more women into the informatics field. The survey results revealed important career development issues for women and roles of professional organizations in cultivating a culture of diversity and inclusivity, providing professional skills training, building networking and connections, improving communications between organizations and members, and across groups within professional organizations. Organized by the same committee, the proposed panel presents the survey findings, and seeks to exchange ideas and suggest solutions targeting female professionals in biomedical informatics for supporting their career development.

Moderator: Deepti Pandita, MD, FACP, FAMIA, is the Chief Health Information Officer at Hennepin Healthcare in Minneapolis, MN. Dr. Pandita is Board Certified in Internal Medicine and Clinical Informatics. She is also Program Director of the Clinical Informatics Fellowship at Hennepin Healthcare. At a national level, Dr. Pandita is active in AMIA Primary care, WAMIA, CICOP and CDS committees and the American College of Physicians (ACP) leading several committees including the ACP Medical Informatics and Telehealth Committee.

Panelist #1: Gretchen Purcell Jackson, MD, PhD, FACS, FACMI, FAMIA, AMIA Board Chair-Elect 2021, IBM Watson Health. Dr. Jackson is Vice President, Chief Health and Science Officer at IBM Watson Health and Associate Professor of Surgery, Pediatrics, and Biomedical Informatics at the Vanderbilt University Medical Center. She is an internationally recognized informatician and accomplished clinical surgeon with over 25 years of...
contributions to informatics research, innovations in health information technologies, and surgical science. Dr. Jackson serves on several journal editorial boards, including Journal of Biomedical Informatics, International Journal of Medical Informatics, and an associate editor of Applied Clinical Informatics and JAMIA Open. Her current research focuses on evaluating clinical decision support systems and empowering patients and families through health information technologies. As the AMIA Board Chair, Dr. Jackson will share her ideas on how a professional organization can support her members’ career development, especially for the female members.

Panelist #2: Omolola Ogunyemi, PhD, FACMI, Director, Center for Biomedical Informatics (CBI), Professor in the Department of Preventive and Social Medicine, Charles R. Drew University of Medicine and Science and co-Chair of the UCLA CTSI’s biomedical informatics program. Dr. Ogunyemi’s research interests include computerized medical decision support, reasoning under uncertainty, 3D graphics and visualization, and machine learning. She has been awarded several NIH grants focusing on informatics solutions for medically underserved communities. Dr. Ogunyemi is an editorial board member of the Journal of Biomedical Informatics. She served as a member of AMIA’s Doctoral Dissertation Award Committee (2017-2020), is an elected Fellow of ACMI and a Member-at-Large of the ACMI Executive Committee. As the current Chair of the WIA Steering Committee, she will share her experience of becoming a leader by collaborating with professionals to contribute to professional organizations.

Panelist #3: William Hersh, MD, FACMI, FAMIA, FACP is Professor and Chair of the Department of Medical Informatics & Clinical Epidemiology in the School of Medicine at Oregon Health & Science University (OHSU). Dr. Hersh is a leader and innovator in biomedical informatics both in education and research. He serves as Director of the OHSU Biomedical Informatics Graduate Program. Dr. Hersh's research focuses on information retrieval and he has authored over 200 scientific papers. Dr. Hersh has won numerous awards for his innovations, including the 2008 AMIA Donald A.B. Lindberg Award for Innovation in Informatics; the Modern Healthcare Top 25 Clinical Informaticists (three times); and the HIMSS Physician Leadership Award. He was elected President of the International Academy of Health Sciences Informatics in 2020. As a senior informatics expert, Dr. Hersh will discuss the presented topics from a male’s perspective, including as a department chair mentoring junior women faculty in their career development.

Panelist #4: Kelly Taylor, Program Manager, Member Services and Engagement, AMIA. Ms. Taylor manages the AMIA Connect platform and she is the membership department’s liaison to all working groups, the WIA steering committee and subcommittees. She manages AMIA’s First Look Program. Ms. Taylor will share her experience in communications within and among professional organizations.

The panel will cover the following topics:
1. Cultivating a culture of Diversity, Equity & Inclusion (DE&I). DE&I related efforts have a long history in the USA. This important topic has gained much more spark and attention in the year 2020, not only in the US but also worldwide. The existing gaps in DE&I efforts and suboptimal outcomes have spanned across various disciplines and professions, and the field of informatics is not an exception. Upholding DE&I core values, AMIA with more than 5000 members from diverse backgrounds has been striving and taking actions to help alleviate the existing gaps in this important area. The panel will discuss how the informatics community, especially a professional organization such as AMIA can foster and nurture a culture of DE&I by advising and guiding, and help executing strategic goals and objectives centered around DE&I activities.
2. Professional development to improve professional qualities and interpersonal skills. The survey results showed that communication, critical thinking, leadership and interpersonal skills are four top skills for career development, especially for female informaticists. Although both female leadership and recognition have been improving in general, a recent publication on gender representation in U.S. biomedical informatics leadership and recognition found that “as in other STEM fields, leadership and recognition in biomedical informatics is lower for women.” These findings inform data-driven strategies to foster diversity and inclusion. The panelists will discuss possibilities for various professional development formats, including providing a second term of WIA Leadership Program, workshops or regular webinars in professional skills trainings and so on.
3. Networking and connection building. Engaging in networking behavior is considered an important career management strategy. Majority of jobs are filled through networking, which means this activity ultimately
benefits us in the long term. However, such study indicates that women’s networks are less powerful than men’s and women are less able to utilize the networks they have due to various reasons, e.g. lack of confidence, missing out on networking opportunities due to childcare responsibilities. For women in informatics, it is crucial to discuss how to build and strategically maintain a network in the biomedical informatics career settings.

4. Effective communications and active publicity from professional organizations. In an era of excessive marketing and consumer avoidance, no other topic could be more interesting than engagement. The value of publicity is dependent on the outcome it produces through different types of media channels. The survey distributed by the committee found that about 50% of respondents never heard of or participated in events organized by the committee. This is a dilemma faced by many non-for-profit organizations, and a strategy to incorporate members' consistent feedback and increase their involvement is compelling to the success of activity publicity in a professional organization. The panel will discuss channels and strategies enhance the communications within and across organizations and groups.

**Significance of the Topic and Anticipated Audience**

The topic of this panel is timely, urgent, needed, or attention-grabbing given the need to:

- Develop a better understanding of career development issues at different stages of women’s lives and careers in biomedical informatics.
- Identify concerns, and develop programs and solutions to provide a supporting culture and environment to foster the career advancement of health informatics professionals, especially women.

This session will be of interest to biomedical informatics professionals at any career stages, especially women, who are interested in learning and preparing for career development issues they may encounter. In addition, professional organizations, such as AMIA, working with their members as a community, may develop policies and programs to provide additional supports to execute strategic plans and fulfill the needs of their members.

**Expected Discussions**

After introductory presentations, the moderator will ask questions and solicit questions from the audience to prompt discussions among the panelists. Potential questions include:

- What resources are currently available to female informaticians that our audience may not be aware of?
- How can we bridge the gap between the resource availability and unawareness or inaccessibility?
- What is a path forward to create Equity and Inclusiveness for women in Informatics, especially addressing current gaps around the representation of women of color?
- What can be done in general to promote career development for women in the informatics field?

**Participation Statement**

All proposed panelists and the moderator are aware of this panel submission, and have agreed to participate in the panel if the proposal is accepted (as of 3/10/2021).

**References**

Public Health Informatics in a Global Pandemic: In the COVID response trenches

Jessica D. Tenenbaum, PhD1,2, Theresa R. Cullen, MD3,8, Neil Sarkar, PhD, MLIS4,5, Philip R.O. Payne, PhD6, Brian E. Dixon, PhD7,8

1Duke University, Durham, NC; 2North Carolina Department of Health and Human Services; 3Pima County Department of Public Health; 4Rhode Island Quality Institute, Providence RI; 5Brown University, Providence, RI; 6Washington University in St. Louis School of Medicine; 7Indiana University Fairbanks School of Public Health, Indianapolis, IN; 8Regenstrief Institute, Indianapolis, IN

Abstract

The COVID-19 pandemic unfolded rapidly around the globe over the course of 2020 - 2021. As the pandemic has progressed, public health agencies and healthcare provider organizations (HPOs) have sought to gather, manage, analyze, and share data and information to track disease spread, establish response plans, and keep the public informed. These responses require informatics methods and tools, such as data integration and harmonization, analytical modeling, data visualization, and a variety of decision support modalities. However, many stakeholders struggled to respond due to inefficiencies, antiquated infrastructure, and poor communications systems. In this panel we will discuss public health informatics related to COVID-19 from different vantage points: state government, regional health information exchange, county government, and HPO partnerships with local and state officials. We will discuss lessons learned, best practices, and what the world and the field should be investing in now to make sure we’re better prepared for the next public health emergency, global or otherwise.

Introduction

Public health informatics, one of the major branches of the field of biomedical informatics, is a discipline in which methodologies are developed and systematically applied to leverage digital health data to support public health practice, research, and learning. Informatics leaders in public health organizations and HPOs that are engaged in public or population health initiatives support policy decision-making with data and information on community health and well-being, including non-medical data on social and environmental determinants. Informatics professionals in public health organizations and HPOs leverage information systems to support the development, implementation, and evaluation of population level interventions that aim to improve the health and well-being of communities. In the context of a pandemic like COVID-19, these interventions include widespread testing, case investigation, contact tracing, exposure notification, isolation support services, hospital capacity planning, and most recently, vaccination campaigns.

Topic importance and timeliness

Beginning in 2020, and continuing into 2021, the COVID-19 pandemic challenged public health systems and HPOs around the world. Data and information were critical to public health organizations’ and HPO’s efforts at tracking disease spread (e.g., screening for new cases), deploying public health workers and care providers to affected areas, and implementing policies (e.g., lockdowns, social distancing, masking). Yet many public health organizations struggled to effectively and efficiently capture, store, manage, analyze, and use data and information when tracking COVID-19 and responding to its presence. HPOs engaged in public and population health initiatives in response to COVID-19 have experienced analogous challenges. These public health informatics struggles, successes, and lessons learned from COVID-19 are important to highlight during the 2021 AMIA Annual Symposium. The pandemic has served to shine a spotlight on the historic underinvestment and consequential shortcomings of public health informatics infrastructure nationally, as well as clinical informatics capabilities needed to deliver and coordinate care across and between traditional organizational boundaries. This has resulted in the renewed recognition of need for increased investment in public health digital infrastructure, the important synergies between public health organizations and HPOs, and the need for infrastructure that can support or enable such data-driven collaboration. Examination and discussion of the challenges and opportunities faced will help guide allocation of newfound resources in this space coming from a number of local, state, federal, and private initiatives.

In contrast with other sessions on the 2021 agenda that may highlight innovations and lessons from the perspective of clinical organizations in their efforts to effectively treat individuals severely impacted by the SARS-CoV-2 virus, this panel will provide knowledge and lessons from the public health perspective.
Speakers:

**Dr. Brian E. Dixon – IU Fairbanks School of Public Health and Regenstrief Institute - Moderator**

Dr. Dixon’s research focuses on applying informatics methods and tools to improve population health in clinical as well as public health organizations. His work leverages clinical and administrative data in electronic health records to measure population health, better understand the determinants of health, examine information flow in the health system, and improve outcomes in individuals and populations. In 2020, Dr. Dixon led a team of informaticians, data analysts, and computer scientists to create an interactive, statewide dashboard on COVID-19 cases, hospitalizations, and mortality for the State of Indiana. The dashboard was used by state and local public health leaders to identify trends, examine sub-populations (especially minority populations), and plan strategies for testing as well as vaccination. Last year he also published the 3rd Edition of the *Public Health Informatics and Information Systems* book in the Springer International Health Informatics series.

Dr. Dixon will introduce the topic of the panel and the panelists. He will then moderate audience questions and discussion after the panelists each present a brief summary of their work during the pandemic.

**Dr. Jessica D. Tenenbaum- NC DHHS; Duke University**

Dr. Tenenbaum has served as Chief Data Officer for North Carolina’s Department of Health and Human Services since 2019 when the NCDHHS Data Office was created. She is also an Assistant Professor of Translational Biomedical Informatics at Duke University. Under Dr. Tenenbaum’s leadership, the NCDHHS Data Office focuses on data democratization, governance, and quality improvement, infrastructure modernization, and analytic innovation. The intent is to facilitate data-driven policy and a “learning DHHS” analogous to the Learning Health System.

In this panel, Dr. Tenenbaum will describe numerous data challenges that COVID-19 has posed for NC DHHS where existing systems were not built to handle the types and volumes of data needs that have arisen in the COVID-19 context. She will describe the early days of the pandemic, where testing volumes were gathered via email. She will also describe the transition of collecting data regarding hospital surge capacity from a process involving manual surveys and pivot charts in Excel to automated HL7 feeds imported into a cloud-based data warehouse, which then feed a public-facing Tableau dashboard. She will describe North Carolina’s ever-expanding public facing COVID-19 data dashboard, including emphasis on downloadable data and demographic stratification. She will highlight both victories and limitations of efforts to better integrate clinical data from NC HealthConnex, North Carolina’s Health Information Exchange (HIE) with public health surveillance data, and of predictive modeling to anticipate demand on hospital capacity. Finally, Dr. Tenenbaum will highlight how data visualization helped inform on-the-ground tactics to improve equity in the push to vaccinate, ensuring allocation and administration of vaccines to portions of the population that have been hardest hit by this cruel pandemic.

**Dr. Neil Sarkar- Rhode Island Quality Institute (RIQI) and Brown University**

Dr. Sarkar is the President and Chief Executive Officer of the Rhode Island Quality Institute (RIQI), which serves as Rhode Island’s Regional Health Information Organization. He is also an Associate Professor of Medical Science and Associate Professor of Health Services, Policy & Practice at Brown University. Prior to his current role at RIQI, he was the founding director of the Brown Center for Biomedical Informatics. Dr. Sarkar’s work is focused on the development of a systemic framework that bridges digital health data to support patients, providers, and payers in achieving the vision of a continuously learning health system.

Dr. Sarkar will share his perspective as a Regional Health Information Organization (RHIO) in supporting both public health and provider community needs. He will describe a RHIO’s role from the early days of the COVID-19 pandemic, and how informatics approaches were developed in near real-time to address myriad information needs, including access to SARS-CoV-2 test results done across multiple laboratory sites and rapid point of care testing, working with community organizations to identify barriers to testing in underserved and high-risk populations in Rhode Island, modeling of health data to support targeted vaccination efforts, developing and implementing approaches to identify putative adverse events in high risk populations that were not included in vaccine clinical trials, and supporting the monitoring of short-term and long-term health impacts of COVID-19 and vaccination. A unifying theme in Dr. Sarkar’s presentation will be an exposition of informatics approaches that can be used to support the pragmatic design and development of sustainable infrastructure for supporting public health needs, which extend beyond pandemic response.

**Dr. Philip Payne, Washington University in St. Louis School of Medicine**

Dr. Payne is the Associate Dean for Health Information and Data Science and Chief Data Scientist at the Washington University School of Medicine. In addition, he is the Janet and Bernard Becker Professor and founding Director of the school’s Institute for Informatics (I2). Dr. Payne is the author of over 200 publications. His current research portfolio focuses on: 1) machine
learning and cognitive computing approaches to the discovery and analysis of bio-molecular and clinical phenotypes; 2) interventional approaches to the use of electronic health records and clinical decision support systems; and 3) the design and evaluation of open-science platforms that enable collaborative and cumulative approaches to scientific discovery.

Since March of 2020, Dr. Payne has led the joint COVID-19 analytics workgroup convened by Washington University and its partners at BJC Healthcare. This workgroup has served to not only support the operational needs of Washington University Physicians and BJC Healthcare in meeting the challenges posed by COVID-19, but has also served as the primary data, information, and knowledge “engine” for the St. Louis Metropolitan Pandemic Task Force, a coalition of elected officials, public health organizations, and HPOs that collectively serve the 3.4M citizens of the St. Louis MSA. Via this unique and collaborative structure, Dr. Payne’s team has rapidly implemented and supported: 1) a regional data sharing network that enables hospital capacity planning and epidemiological surveillance of COVID-19 cases; 2) comprehensive case management, contact tracing, and exposure notification solutions that have facilitated the monitoring and containment activities of five different public health organizations serving the region as well as major employers and universities; 3) data management and reporting platforms that allow for the delivery of an inexpensive and rapid-cycle saliva-based COVID-19 testing assay developed by Washington University for use by multiple regional partners; and 4) pre-registration, tracking, and scheduling technologies that support public vaccination campaigns spanning high-throughput HPOs and public health organizations. In all of the aforementioned cases, the work of Dr. Payne’s team has served to demonstrate the unique opportunities afforded by substantive partnerships between HPOs and public health organizations to offset infrastructure issues and ultimately respond to the public health needs of the region, thus identifying new opportunities to build a public health informatics “fabric” that can both scale and be sustained in the long-term.

Dr. Theresa Cullen- Pima County AZ Department of Public Health; Regenstrief Institute

Dr. Cullen is currently the Public Health Director at Pima County, Arizona, a county of over 1 million people that encompasses urban and rural communities as well as two American Indian reservations. In this role, she has overseen the Pima County COVID-19 response, including testing, case investigation, contact tracing and the accelerated immunization plan. Pima County surpassed over 120K cases of COVID 19 in the first 12 months of the pandemic, and also delivered over 300K immunizations in January and February 2021.

Dr. Cullen will talk about the epidemiological cascade and data needs from a County perspective. In addition, she will share insights about the vaccine delivery system, and the impact of limited public health infrastructure on health equity at a county level. Pima County is committed to health justice and equity, and successfully supported novel technologies as well as interventions to accelerate a healthy community during the COVID epidemic. However, there continue to be multiple public health informatics opportunities that will need to be addressed in the future, including the need for rapid reconciliation of patient identification, near real time longitudinal sharing of infectious disease test results and demographics with local as well as state based public health agencies for early sentinel awareness as well as intervention, the minimum data set for rapid registration and longitudinal tracking at mobile vaccination sites, and ongoing efforts to achieve equity throughout the epidemiological, therapeutic and immunization cascade.

Questions to be addressed:

- What were the major obstacles in the COVID-19 pandemic response from a data/informatics perspective?
- What data/informatics lessons were learned by public health agencies and collaborating HPOs as their response efforts unfolded?
- What were the greatest successes or opportunities for innovation and impact?
- What technological/infrastructure/data challenges remain to be addressed post-pandemic?
- What unique cross-sector partnerships are needed to address these technological/infrastructure/data challenges moving forward?

Conclusion

This collection of panelists represents a diverse set of perspectives and approaches to this timely topic. The topic could not be timelier and addresses the public health end of the informatics spectrum that has been top of mind throughout the pandemic, not only for our field but for the world.
Leveraging PGHD in Clinical Workflows: Opportunities and Challenges for Use in Patient Care

Victoria L Tiase PhD, RN-BC, FAAN, FAMIA1,2; Robin R Austin PhD, DNP, DC, RN-BC, FAMIA3; Christie L Martin MN, MPH, RN, PHN, LHIT-HP3; Young Ji Lee PhD, RN4;

1Department of Information Technology, New York-Presbyterian Hospital, New York, NY; 2College of Nursing, University of Utah, Salt Lake City, UT; 3School of Nursing, University of Minnesota, Minneapolis, MN; 4University of Pittsburgh, Pittsburgh, PA

Abstract
Given the emphasis on digital health solutions and the increasing number of mobile health applications, patient-generated health data (PGHD) are abundant. However, without the availability of PGHD at the point of care, some of their utility may be lost. If presented to clinicians at the point of care, PGHD have the potential to improve patient outcomes by providing new insights into the patient’s longitudinal status and decreasing nursing documentation burden. This interactive panel of nursing informatics leaders will discuss key findings from several studies aimed at leveraging PGHD within clinical workflows. Session participants will have the opportunity to discuss successes and challenges, suggest a road map for additional research, and envision PGHD use cases across the continuum of care.

General Description
Patient generated health data (PGHD) are data created, recorded, or gathered by or from patients (or family members or other caregivers) using smartphones, sensors, or other medical devices.1 PGHD can be incorporated into clinical practice at the point of care and may facilitate communication between clinicians and patients. Longitudinal data, for example, provide insights into patients’ health behaviors and outcomes. In tandem with their patients, clinicians are well suited to interpret these data and find patient-centered solutions to improve overall health and well-being. Specifically, using their assessment skills, nurses are instrumental in advancing care practices that utilize PGHD.

Although PGHD can provide value to the patient’s overall health status, there are very few examples of their use within the clinical setting. In 2019, a systematic review explored the impact of PGHD in clinical settings and found just 21 studies, including limited integration of PGHD into electronic records.2 A scoping review, published in 2020, reported only 19 studies that described the characteristics of PGHD integration to electronic health records (EHRs) and found that efforts are needed to optimize the incorporation of PGHD into clinical workflows.3 Furthermore, there is concern that adding more information to an EHR that contains a large amount of outdated regulatory requirements may contribute to provider burnout.4 Additional concerns surrounding the use of PGHD in clinical practice include time to review data, scope of practice, liability for data that is received, and appropriate interpretation and clinical action based on the data. PGHD may potentially be lost, excluded, or hidden due to a lack of standards and are often incorporated in the EHR as unstructured data.

Incorporating PGHD into workflows, both technically and operationally will require careful consideration of information needs that support clinical decision making. Data standards, including standardized terminologies (i.e., SNOMED CT and others), can provide structure to the data facilitating interoperability, thereby enabling patients to share their voice within the EHR. This interactive panel includes a variety of PGHD integration topics based on the panel participants’ collective work. The panel aims to engage a wide variety of participants (clinicians, researchers, CMIOs, CNIOs, etc.). Panelists will discuss approaches to the design and implementation of mobile health applications as well as care interventions that encourage the collection, use, and sharing of PGHD.
Discussion Topics

Panel members will present their recent findings, personal perspectives, institutional experience, and plans to engage audience members in a robust discussion to elicit participants’ opinions, experiences, and practices with PGHD use.

The following topics will be addressed:

• The role of the nurse in supporting patients and families with their PGHD ○ Highlighting novel COVID-19 use cases
• The impact of PGHD on patient outcomes ○ Including opportunities for new care models and community outreach
• Interoperability and standards for PGHD ○ Research related to visualization and summarization
• Types of PGHD that may be of interest to clinicians ○ Examples using pain scale data, environmental data, social media

Along with eliciting questions from the audience, we will use the following questions to guide the discussion:

1. What challenges have you experienced integrating PGHD into clinical practice?
2. Describe successful PGHD use cases that you have found to be beneficial?
3. What specific types of PGHD are best suited to be shared by patients (families) and clinicians?
4. What areas of PGHD research require further exploration?

Panel Members

Victoria L Tiase, PhD, RN-BC, FAAN, FAMIA is a nurse, informatician, and Director of Research Science at NewYork-Presbyterian Hospital. She supports a range of clinical information technology projects related to patient engagement and is passionate about the integration of patient generated health data into clinical workflows. She serves on the steering committee for the Alliance for Nursing Informatics and is a board member of AMIA, NODE.Enh, and the CARIN Alliance. Recently, she was appointed to the National Academy of Medicine Committee on the Future of Nursing 2030. In addition, Dr. Tiase is an Assistant Professor in the Division of Health Informatics at Weill Cornell Medicine. She completed her BSN at the University of Virginia, MSN at Columbia University, and PhD from the University of Utah with a focus on the integration of PGHD into clinical workflows to impact care and reduce documentation burden. Dr. Tiase, also serving as moderator, will present her work on provider preferences for the graphical display of pediatric-asthma PGHD to support decisions and information needs in the outpatient setting. She will also discuss the need for future research that examines the interactive nature of PGHD and EHR data.

Robin R Austin, PhD, DNP, DC, RN-BC, FAMIA is an Assistant Professor at the University of Minnesota, School of Nursing. Dr. Austin has over 20 years of clinical healthcare experience and has translated this experience to focus on patient-centered research. Dr. Austin has earned a Doctorate in Nursing Practice (DNP) and PhD in Nursing, both specializing in Nursing Informatics. Dr. Austin integrates her clinical background with informatics methods to represent the patients’ perspectives across the healthcare continuum. Through her research, Dr. Austin seeks to empower individuals through the use of technology and include their voice in person-centered care. Dr. Austin will present her research examining PGHD from a web-based application, entitled MyStrengths+MyHealth, from a virtual community outreach initiative examining individual and community strengths, challenges, and needs during COVID-19.

Christie L Martin, MN, MPH, RN, PHN, LHIT-HP is a medical-surgical nurse at Abbott Northwestern Hospital in Minneapolis and a Research Assistant for the Child and Family Health Co-
Operative Unit at the University of Minnesota, School of Nursing. Ms. Martin is a doctoral candidate at the University of Minnesota School of Nursing. Her research focuses on child and adolescent health promotion and informatics with an emphasis in behavior change and mobile health technologies. She completed a Master of Nursing and a certificate in Leadership in Health Information Technology for Health Professionals from the University of Minnesota, School of Nursing and has a Master of Public Health from the University of Minnesota, School of Public Health. Ms. Martin is the outgoing co-chair of the Nursing Knowledge Big Data Science mHealth for Nursing working group and the VP of Membership of the AMIA Consumer and Pervasive Health Informatics Working Group. **Ms. Martin will present findings of a systematic review investigating the efficacy and components of pain-related mHealth apps and speak specifically to the sharing of PGHD between consumers and providers of eligible studies.**

**Young Ji Lee PhD, RN,** is an Assistant Professor at the University of Pittsburgh School of Nursing and Department of Biomedical Informatics. Dr. Lee has focused on structuring and delivering health information through an informatics-based approach for diverse groups and underserved populations. She has completed several projects on PGHD collected from online cancer forums to understand the unmet needs of cancer patients and caregivers. Currently, Dr. Lee is a PI of the NIH-funded study to build and implement a personalized information access system with a hybrid recommender engine that adapts to different patient needs: HELPeR—Health E-Librarian with Personalized Recommendations. Currently, she is serving on the program committee for the AMIA 2021 Informatics Summit and AMIA 2021 Clinical Informatics Conference. Dr. Lee was elected as a member-at-large of the AMIA Nursing Informatics Working Group. She is also a member of the ANI Emerging Leaders Program which aims to understand facilitators and barriers of PGHD implementation in clinical settings. **Dr. Lee will discuss nurses’ perceived challenges of PGHD implementation in the clinical settings. She will also present how researchers and clinicians can utilize PGHD collected from social media to understand patients’ needs and perspectives.**

**References**

Lessons in Health Equity from the CMS AI Health Outcomes Challenge

David K. Vawdrey PhD, FACMI¹, Dave DeCaprio², Carol McCall FSA, MPH³, Yaron Kinar PhD³, Aneesh Chopra, MPP⁴

¹Steele Institute for Health Innovation, Geisinger, Danville, PA
²Closedloop.ai, Austin, TX
³Medial Earlysign, Tel Aviv, Israel
⁴CareJourney, Arlington, VA

Abstract

In April 2021, CMS announced the winner and runner-up of the CMS Artificial Intelligence Health Outcomes Challenge, the largest-ever prize competition for innovators to demonstrate how AI solutions can predict patient health events. The two-year competition was described as “an exciting example of how public/private partnerships can drive innovation.” This panel will explore how the winner and runner-up in the AI Challenge approached health equity by measuring and mitigating algorithmic bias. The panelists will discuss the current and future state of algorithmic bias in applications using healthcare claims data, debating topics such as: 1) how to effectively identify the sources of bias in AI algorithms—including data quality and representativeness, 2) the difference between bias and fairness, 3) the value of scoring systems for quantifying bias, and 4) practical implications of applying predictive models in healthcare compared to other settings. Ample time will be provided for audience questions and discussion.

General Description

The CMS Artificial Intelligence Health Outcomes Challenge launched in 2019 with more than 300 entrants representing the world’s leading technology, healthcare, and pharmaceutical innovators. Entrants competed to create explainable AI solutions that clinicians could trust to predict health outcomes, target scarce resources, and keep their patients healthy. Sponsored by CMS’s Center for Medicare and Medicaid Innovation (CMS Innovation Center) in collaboration with the American Academy of Family Physicians and Arnold Ventures, the AI Challenge was designed to accelerate development of AI solutions for predicting patient health outcomes for Medicare beneficiaries for potential use by the Innovation Center. With prizes totaling $1.6M, the AI Challenge is the largest healthcare AI competition to date. The winner, announced in April 2021, was ClosedLoop.ai (Austin, TX), with Geisinger (Danville, PA) in partnership with Medial Earlysign (Tel Aviv, Israel) as runner-up.

Health equity was at the heart of CMS AI Challenge. In addition to industry-standard measures of algorithmic accuracy (e.g., area under the ROC curve, calibration, and precision) and data visualization (to explain predictive models’ output to clinicians and patients in an understandable, useful, and trustworthy way), the AI Challenge judged finalists on their ability to address the issue of implicit algorithmic bias in their submissions. ClosedLoop.ai and Geisinger/Medial Earlysign proposed creative approaches to mitigating potential bias.

This panel will educate healthcare providers and informatics professionals about the challenges associated with algorithmic bias in predictive models based on healthcare claims data. The panel participants represent the winner and runner-up in the CMS AI Challenge, and they have extensive experience developing and implementing health information technology in a just and equitable manner. Panelists will synthesize their perspectives on the biomedical literature and future developments in this area. They will explore the following discussion topics and engage in stimulating dialog based on questions from the audience.

Discussion Topics

How to effectively identify the sources of bias in AI algorithms

Representativeness and data quality are two issues to consider in terms of their impact on algorithmic bias. When discussing representativeness, a question that algorithm implementers may ask is, “To what extent did the data used to train the algorithm come from a population that is representative of the intended recipients?” In the CMS AI Challenge, through a data use agreement, teams were provided with a 5% sample of the CMS Limited Data Set (LDS) Files, containing beneficiary level health information (such as claims) stripped of specified direct identifiers [1]. The Table compares age, race/ethnicity, and sex from the CMS LDS sample with data from the entire

160
U.S. population obtained from the U.S. Census Bureau. Beneficiaries in the LDS data are much older on average than the U.S. population, include more females, and considerably underrepresent racial and ethnic minorities.

Even if a training data set is “representative” of the intended delivery population, data quality remains an issue. Polubriaginof and colleagues reported that among the 160 million combined patients from two large observational health data sets (one public and one private), race or ethnicity was unknown for 25% [2]. In the CMS Limited Data Set, for example, race and ethnicity are conflated into one construct; “Two or more races” is not an option, and options for Sex are limited to “Female” or “Male.” Other important social determinants of health (e.g., primary language, education attainment, household income, veteran status, sexual orientation) are often completely missing, especially in data sets with healthcare claims.

**Bias vs. fairness**

AI platforms should systematically assess for bias in model design, data, and sampling, and they should use measures such as Michael’s Correlation Coefficient that are insensitive to differences in disease prevalence between groups. Additionally, it’s important to understand the concept of fairness in relation to bias. Assessing fairness can involve questions such as, “to what extent does the output of the algorithm differ for various groups in a manner that is not clinically justifiable and/or is ethically tenuous?” [3].

In the AI Challenge, ClosedLoop.ai developed a new metric for measuring fairness called Group Benefit Equality (GBE). Standard fairness metrics (e.g., Disparate Impact) are not well suited to healthcare situations. They ignore false-negative errors, which can leave individuals who would benefit from an intervention unable to get it, or they use arbitrary benchmark thresholds that fail to adjust for instances where the alarm rate for the reference group is too low. The GBE metric addresses these shortcomings. It is also easily explained, has transparent procedures, and uses clearly defined thresholds to assess when models are biased [4]. The Geisinger/Earlysign team applied the simple fairness criterion suggested by Hardt et al. [5] which helps evaluate the equality of opportunities for various subgroups. It is important to understand the root cause of algorithmic bias, which may arise from inherent differences between subgroups in demographic characteristics (e.g., age or sex distribution). Unfortunately, mitigation strategies often involve a tradeoff between fairness and model performance—which should lead to a lively discussion amongst panelists and audience members.

**The value of “scoring systems” for quantifying bias**

The underlying data used to train many healthcare AI models—including the CMS data using in the AI Challenge—reflects historical biases and inequities in access and utilization. The increasing reliance on algorithms to do things like target interventions, reward performance, and distribute resources has put the notion of quantifying algorithmic bias and fairness in the spotlight. This is particularly important in settings where AI models could be used to help prioritize the allocation of limited resources—for example, to decide which Medicare beneficiaries a care manager should reach out to, based on the beneficiaries’ algorithmically-determined risk of an unplanned hospital admission.

Models that systematically underpredict risk for a particular group can lead to that group being unfairly denied resources. For example, the recent Algorithmic Bias Playbook developed by Obermeyer and colleagues based on their previous work [6,7] describes how algorithm implementers should be cautious of conflating healthcare costs (i.e., utilization) with healthcare needs. Even though cost and health needs are correlated, two patients with the same level of need might have considerably different costs if one receives less care. The authors poset that we should not necessarily blame algorithms when their predictions “reinforce and perpetuate systemic racism,” for example, when an algorithm that was trained to predict healthcare costs (which are historically higher for white vs. Black Americans) is improperly applied to as a proxy for healthcare needs (and Black individuals receive less care as a result). Such findings must be considered when evaluating predictive models for bias based on race, ethnicity, sex, age, disability status, and other characteristics.

**Table.** Demographics of CMS Limited Data Set vs. U.S. Population.

<table>
<thead>
<tr>
<th>Category</th>
<th>CMS Limited Data Set</th>
<th>U.S. Population</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sample size</strong></td>
<td>2,846,420</td>
<td>328,239,523</td>
</tr>
<tr>
<td><strong>Age</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0-17</td>
<td>0.04%</td>
<td>22.3%</td>
</tr>
<tr>
<td>18-64</td>
<td>34%</td>
<td>61.2%</td>
</tr>
<tr>
<td>65+</td>
<td>66%</td>
<td>16.5%</td>
</tr>
<tr>
<td><strong>Race/Ethnicity</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>American Indian or Alaska Native</td>
<td>0.5%</td>
<td>1.3%</td>
</tr>
<tr>
<td>Asian</td>
<td>2.5%</td>
<td>5.9%</td>
</tr>
<tr>
<td>Black or African American</td>
<td>10.5%</td>
<td>13.4%</td>
</tr>
<tr>
<td>Native Hawaiian or Other Pacific Islander</td>
<td>-</td>
<td>0.2%</td>
</tr>
<tr>
<td>Two or more races</td>
<td>-</td>
<td>2.9%</td>
</tr>
<tr>
<td>White</td>
<td>80%</td>
<td>76.3%</td>
</tr>
<tr>
<td>Hispanic or Latino</td>
<td>3%</td>
<td>18.5% [1]</td>
</tr>
<tr>
<td>Other/Unknown</td>
<td>3.5%</td>
<td>-</td>
</tr>
<tr>
<td><strong>Sex</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>54%</td>
<td>50.8%</td>
</tr>
<tr>
<td>Male</td>
<td>46%</td>
<td>49.2%</td>
</tr>
</tbody>
</table>

[1] The value of “scoring systems” for quantifying bias. The underlying data used to train many healthcare AI models—including the CMS data using in the AI Challenge—reflects historical biases and inequities in access and utilization. The increasing reliance on algorithms to do things like target interventions, reward performance, and distribute resources has put the notion of quantifying algorithmic bias and fairness in the spotlight. This is particularly important in settings where AI models could be used to help prioritize the allocation of limited resources—for example, to decide which Medicare beneficiaries a care manager should reach out to, based on the beneficiaries’ algorithmically-determined risk of an unplanned hospital admission.

Models that systematically underpredict risk for a particular group can lead to that group being unfairly denied resources. For example, the recent Algorithmic Bias Playbook developed by Obermeyer and colleagues based on their previous work [6,7] describes how algorithm implementers should be cautious of conflating healthcare costs (i.e., utilization) with healthcare needs. Even though cost and health needs are correlated, two patients with the same level of need might have considerably different costs if one receives less care. The authors poset that we should not necessarily blame algorithms when their predictions “reinforce and perpetuate systemic racism,” for example, when an algorithm that was trained to predict healthcare costs (which are historically higher for white vs. Black Americans) is improperly applied to as a proxy for healthcare needs (and Black individuals receive less care as a result). Such findings must be considered when evaluating predictive models for bias based on race, ethnicity, sex, age, disability status, and other characteristics.
Practical implications of applying predictive models in healthcare compared to other settings

One well-known application of predictive algorithms outside of healthcare is the Netflix Recommender System [8]. Similar systems exist to connect individuals (e.g., via Facebook or LinkedIn), to target advertising, or to suggest products to prospective buyers. Predictive models in healthcare may have much in common with these applications, but there are also important differences. As Halamka and Cerrato recently explained, when Netflix “recommends a new movie on the basis of a person’s previous movie choices, the viewer may find the suggestion helpful, annoying, or even amusing — but not life-threatening. When algorithms are used to assist in the diagnosis of eye disease, melanoma, or sepsis, the stakes are obviously much higher because misleading advice can have far more serious consequences” [9].

Besides the life-threatening stakes that exist in healthcare versus other industries, implementing predictive models in the healthcare setting is complicated by ethical questions related to beneficence, by the rapidly evolving regulatory environment and by questions involving healthcare payment models. One idea that touches on these topics is Eaneff and colleagues’ proposal that healthcare organizations adopt Algorithmic Stewardship for AI and machine learning technologies [10].

Panel Members  (All members have agreed to participate on the panel)

David K. Vawdrey, PhD, FACMI is Chief Data & Informatics Officer for Geisinger. His research interests include using health information technology to improve value, patient engagement, and safety.

Dave DeCaprio is Chief Technology Officer and co-founder at ClosedLoop.ai. He has two decades of experience in healthcare and life sciences AI and led the ClosedLoop.ai team in the CMS AI Challenge.

Carol McCall FSA, MPH is Chief Health Analytics Officer at ClosedLoop.ai. She is a health actuary and population health executive whose specialty is combining analytics, health service innovations, and business models in ways that catalyze value-based healthcare.

Yaron Kinar, PhD is Chief Scientific Officer at Medial EarlySign, and he co-led the Geisinger/Medial EarlySign data science team in the CMS AI Challenge. He has 20 years of experience in AI/ML in health-related domains.

Aneesh Chopra, MPP will moderate the panel. He is President of CareJourney, an open data service that helps providers, payers and pharma market leaders make smarter decisions in the move to value. He served as the first U.S. Chief Technology Officer (2009-12) where he spearheaded efforts to make CMS data available for commercial use.

References

Opportunities and Challenges in Health and Clinical Informatics Careers and the Future of the Profession

Sponsored by the Continuing Professional Development Committee and Physicians in AMIA

Amy Y. Wang, MD, MBI1, Jodi Kodish-Wachs, MD2, Deepti Pandita, MD3, Thomas Agresta, MD, MBI4, Catherine H. Ivory, PhD, RN-BC, RNC-OB, NEA-BC5
1University of Alabama at Birmingham, Birmingham, AL; 2Vizient, Chicago, IL; 3Hennepin Healthcare, Minneapolis, MN; 4University of Connecticut, Farmington, CT; 5Vanderbilt University, Nashville, TN

Abstract

Health professionals who practice health informatics (HI) and clinical informatics (CI) are a core constituency in the AMIA community. This diverse group represents medicine, nursing, dentistry, pharmacy, public health, and other health disciplines. There are new developments, opportunities, and challenges as HI and CI evolves, including training, certification, continuing professional development, career paths and stages, and diversity, equity, and inclusion (DEI). In this interactive panel, HI and CI professionals representing diverse fields and stakeholder groups will discuss topics that are current and relevant to the HI and CI community. This panel is co-sponsored by the Continuing Professional Development Committee (CPD, formerly MOC Committee) and Physicians in AMIA (PINA, formerly CICOP) and supports AMIA’s strategic directive of promoting and building the field of informatics.

Learning Objectives

1. Discuss the current landscape and new developments in health and clinical informatics careers, including core content, competencies, training, certification, and continuing professional development (CPD).
2. Discuss current and future challenges and opportunities in health and clinical informatics careers and certification, including promoting diversity, equity, and inclusion (DEI).
3. Evaluate and plan various career paths and opportunities for learning, training, certification, and practice.

Introduction

Healthcare professionals in health informatics (HI) and clinical informatics (CI) apply informatics principles to improve health outcomes through the use of health information. While HI and CI may differ, there is significant overlap. CI refers to the field in which physicians become certified by a member board of the American Board of Medical Specialties (ABMS), while HI refers to the field in which health professionals become certified through AMIA’s Advanced Health Informatics Certification (AHIC). Excluding certification, for practical purposes, HI and CI are often grouped together and even used interchangeably. We advocate considering these fields together when possible, given the similarities and the benefits of a united community.

In response to the rapid growth of HI and CI and the need for specialists in the field, AMIA community has defined the CI specialty, core content, and training programs1-3. Initial efforts focused on board-certified physicians and have expanded to other eligible health professionals4-5. There has been ongoing research and discussion about the scope of practice, workforce, training programs, and board certification6-8. Given the wide scope of informatics practice in healthcare and diverse backgrounds of professionals entering the field, training and career paths ideally are individualized rather than uniform. Career paths may include varying combinations of leadership, administration, management, operations, quality improvement, teaching, research, software design, and information management within various settings (e.g., academic, community, industry, government).

Our aim is to present updates and opportunities and challenges in the HI and CI profession. Panelists represent diverse perspectives and will introduce topics and facilitate an interactive discussion. Topics and questions include:

1. **Impact of COVID-19.** (Deepti Pandit) How has COVID-19 and remote working impacted the field? How have training programs been affected? How did numbers and enrollment change? Did training become remote, and
what was the impact? Did trainees need to cross-cover other areas? Was education shortened with competing interests or expanded? How did certification numbers change? Were there unexpected effects?

2. **ABPM certification update.** (Thomas Agresta) What is the latest news from the ABPM, ABPath, and ACGME? How does one get certified? What changes have there been to eligibility for certification? Will the practice pathway be extended beyond 2023? What is longitudinal assessment, what do I need to know about it, and how do I get started? Will that change be permanent, meaning could there be no more high-stakes exams?

3. **Advanced Health Informatics Certification (AHIC) and Nursing Update.** (Catherine Ivory) What is AHIC? Who needs AHIC, and why? What does AHIC offer compared with CI certification? How and when can I qualify, apply, and prepare for the board exam? What certification do you recommend for nursing informatics?

4. **CI Training and certification.** (Deepti Pandita) Is board certification in CI necessary? Why and when? Physicians typically complete fellowships early in their careers, before entering practice. Will hiring organizations require board certification? If so, will aspiring mid-career health physicians be willing to leave established practices, start as fellows, and endure disruptions affecting the whole family, such as relocation and lost income? How can CI fellows continue practicing and maintaining their knowledge and skills in their primary specialties?

5. **Standard health and clinical informatics roles, qualifications, and job descriptions.** (Jodi Kodish-Wachs) Why are current job descriptions challenging? Why do roles not translate well to core competencies? HI and CI come from diverse backgrounds. For example, physicians differ in technical expertise, residency and fellowship completion, licensure, board certification, clinical experience, and current practice. How do these differences affect competencies, job descriptions, and roles? What are standard roles and competencies? How roles and competencies be more standard and defined to improve job searches based on keyword searches of roles, job descriptions, and qualifications? How can we crowdsource and share this effort?

6. **Diversity, equity, and inclusion (DEI).** (Deepti Pandita) Research has shown that more diverse groups improve creativity and problems solving. How can we promote DEI in HI and CI to encourage more women and underrepresented minorities to enter informatics and training programs and then thrive in HI and CI careers increase ethnic and gender representation and DEI to develop a cohesive yet diverse informatics workforce?

7. **Continuing professional development (CPD) and continuing certification (MOC).** (Amy Wang) What are the latest developments in CPD, including CE, CME, and MOC? How can I stay current in the profession? What happened to the multiple-choice questions for AMIA activities? Where can I get help with MOC? How do I complete MOC IV when I am not in a healthcare setting or have access to patient data, or I am self-employed?

**Intended Audience**

This panel is intended to include all healthcare professionals, trainees, educators, and mentors who are interested in learning about HI and CI changes, training, careers, and certification. This topic is timely given the role of HI and CI informatics in the COVID-19 pandemic as well as the many recent changes in certification requirements.

**Panelists**

**Amy Y. Wang, MD, MBI, FAAFP, FAMIA** (Organizer and Moderator) is Associate Professor of Medicine and Scientist in the Informatics Institute at UAB. She serves as Chair of the AMIA Continuing Professional Development (CPD) Committee and Member of the Women in AMIA Career Advancement Committee. A family physician, her expertise is in standards and interoperability, health terminology, electronic health records, and informatics education. She completed a medical degree at Northwestern University and a master’s degree in biomedical informatics at OHSU. Dr. Wang’s industry experience includes SNOMED CT development at the College of American Pathologists, EHR development at Greenway Medical, and terminology development at IMO. Most recently, she was Assistant Professor and Faculty Director of the Master of Health Informatics degree program at Northwestern University.

**Jodi Kodish-Wachs, MD, FAMIA** is a physician executive at Vizient and serves as Chair of Physicians in AMIA (PINA). She has eight years of clinical practice and nine years of experience at Cerner and Siemens, bridging people, clinical medicine, and information technology to optimize access to health information and tools, individualized health, and improved efficiency. She has practical expertise in ideation, strategy, prototyping, and validating start-up solutions, research, data analysis, health standards, AI, clinical modeling, knowledge management, population health, risk, outcomes measures and care processes. Dr. Kodish-Wachs graduated from Rutgers Robert Wood Johnson Medical School and trained in Physical Medicine and Rehabilitation at UAB.
Deepti Pandita, MD, FACP, FAMIA is Chief Health Information Officer, Director of the Clinical Informatics Fellowship, and Staff Physician at Hennepin Healthcare as well as Assistant Professor of Medicine at University of Minnesota Medical School. Her areas of expertise are care delivery, innovation, population health, data analytics and clinical informatics, and her passion is using technology to bridge health disparities, especially in Medicaid populations, to transform care and create healthy communities. Dr. Pandita completed her residency in internal medicine at University of North Dakota School of Medicine and Health Sciences. Prior to joining Hennepin, she served as physician EMR champion at Park Nicollet Health Services.

Thomas Agresta, MD, MBI is Professor and Director of Medical Informatics in the Department of Family Medicine at and Director of Clinical Informatics in the Center for Quantitative Medicine at University of Connecticut School of Medicine. He is Vice-Chair of the Clinical Informatics Sub-Board of ABPM. Dr. Agresta is a seasoned family physician, educator, administrator, researcher, and innovator who builds multidisciplinary teams to develop methods for creating, using, and evaluating technology in clinical and teaching settings. He completed a BS in biomedical engineering at Stevens Institute of Technology, medical degree at New Jersey Medical School and master’s degree in biomedical informatics at OHSU. He has held state-level leadership roles for health information exchange initiatives.

Catherine Ivory, PhD, RN-BC, RNC-OB, NEA-BC, FAAN is Senior Director of Nursing Research and Assistant Professor of Nursing at Vanderbilt University Medical Center and Vice-Chair of the AMIA Health Informatics Certification Commission (HICC), which oversees Advanced Health Informatics Certification (AHIC). Dr. Ivory earned her Master of Science in Nursing (MSN) from Georgia College & State University and a PhD in Nursing Science from Vanderbilt University. She practiced inpatient obstetric nursing and is a certified informatics nurse and a health services researcher especially interested in the use and re-use of data generated in nursing practice.

Participation Statement

I, Amy Y. Wang, confirm that all speakers have agreed to participate in this panel. Panelists are aware that no funds are available and that the CPD Committee and PINA are unable to reimburse registration or travel costs.

References

Panel: Demonstrations in Synthetic Data and the National COVID Cohort Collaborative (N3C)

Adam Wilcox PhD,a Randi Foraker PhD MA,b Jason A. Thomas BS,c
Jon D. Morrow MD MA MBA,d,e Noa Zamstein PhDf

aUniversity of Washington School of Medicine, Seattle, WA; bWashington University in St. Louis, School of Medicine, St. Louis, MO; cMDClone Ltd., Be’er Sheva, Israel; dNew York University School of Medicine, New York, NY

General Description

The COVID-19 pandemic has been significant in demonstrating the need for sharing data across institutions for broader analysis. As many organizations were facing different waves of the pandemic at different times in 2020, lessons learned from others experiences could be helpful in identifying indicators of pandemic trends, studying the effect of new therapeutics, and understanding variations due to population factors. Most sharing during the pandemic was via scientific hypotheses or demonstrations, or through public health reporting data. The National COVID Cohort Collaborative (N3C) gathered electronic health record data (EHR) from multiple academic medical centers across the country to allow deeper analysis of shared data. By centralizing the data for the cohort, studies could be performed with advanced analytics comparing multiple institutions and with analysts beyond existing institutional analytics groups. However, expanding access to data across multiple institutions requires new approaches to protecting patient privacy. Standard methods of data de-identification that involve date removal, shifting or obfuscation may not be useful when studying data for a pandemic, when predictions about patient severity may depend on different periods of the pandemic or when trying to use the data for overall incidence prediction. Creating synthetic data is seen as an approach that can both protect patient confidentiality and allow broader sharing of data for analysis and discovery. Because of this opportunity, it is important to demonstrate how synthetic data can be used effectively for analysis.

This panel includes a group of researchers who have studied the use of synthetic data for various real analyses related to the COVID-19 pandemic. Using a synthetic dataset derived from N3C data, they have performed studies of data characterization, epidemic measurement, prevalence prediction, and inpatient severity prediction, and compared the data to results drawn from the original dataset. Two of the individual studies have been submitted for publication. The panel provides a comprehensive review of how synthetic data can be used for analysis. While each of the studies could be considered separately, the combination of studies together creates an opportunity to clearly recognize the status of synthetic data validation, which can better advance the appropriate application of the technology in informatics broadly. Discussions with the audience will both provide clarity of the extent of the analysis and discussions of how synthetic data can best be used in our field going forward. This topic is timely due to the analysis of synthetic COVID data, with directly-relevant conclusions that can be applied to current COVID research. It is also important to discuss in our field as we pursue new models of EHR data sharing in disease areas beyond COVID, in ways that preserve patient privacy.

Panelists

Randi Foraker, PhD, MA, FAHA, FAMIA. Dr. Foraker is an epidemiologist and the Director of the Center for Population Health Informatics at the Institute for Informatics and a Professor of General Medical Sciences at Washington University School of Medicine in St. Louis. Dr. Foraker also serves as Director of the Public Health Data and Training Center for the Institute for Public Health. She specializes in the design of population-based studies and utilizing electronic health record (EHR) data for research. Her recent research has focused on the statistical validation of computationally-derived data comparing original data to synthetic data generated by MDCClone (Be’er Sheva, Israel). In her role on the National COVID Cohort Collaborative’s (N3C) synthetic data task team, she led the evaluation of three use cases comparing original to synthetic data, including (1) exploring the distributions of key features of the COVID-positive cohort; (2) training and testing predictive models for assessing the risk of admission among these patients; and (3) determining geospatial and temporal COVID-related measures and outcomes, and constructing their respective epidemic curves. She will be presenting the analyses exploring distributions of key features among the cohort and predictive models for assessing admission risk.

Jason A. Thomas, PhD Candidate. Mr. Thomas is a 4th year PhD candidate in the Department of Biomedical Informatics and Medical Education at the University of Washington with 10 years of experience in healthcare - both frontline healthcare delivery and analysis/modeling of the resulting data. He has expertise in analysis & data modeling of EHRs, clinical trials and longitudinal cohort data. His research interests are currently focused on data access and sharing, privacy preserving technologies, and data utility & quality. He is a member of the JAMIA editorial board and has had an active role in the N3C Synthetic Validation Task Team since June 2020. He will be discussing his analysis of synthetic geospatial and temporal epidemiologic data utility. Analyzing regular and synthetic data across the country in the N3C dataset, he demonstrated the capability of synthetic data for analyzing
temporal trends in prevalence, hospitalization and death. Analyses of geographic areas were done for individual zip codes; where zip codes represented regions with multiple patients, epidemic curves were similar with synthetic data and actual data. Differences were observed in zip codes with small numbers of cases, which is consistent with expectations due to data censoring to preserve patient privacy in the synthetic data derivation process.

Jon D. Morrow, M.D., M.A., M.B.A., F.A.C.O.G. Dr. Morrow is a medical informaticist and obstetrician-gynecologist who currently serves as the Senior Vice President and Physician Executive for MDClone, responsible for the company’s Medical Affairs activities in North America. As a representative of MDClone on the N3C Synthetic Clinical Data Validation Subgroup, he has been involved in the Subgroup’s efforts to demonstrate the utility of synthetic data for extracting new knowledge about COVID-19 from the vast N3C database. He is also a member of the AMIA Informatics Partnership Council. For this panel, Dr. Morrow will be discussing an analysis of synthetic data across multiple geographic regions used to predict epidemic curves for 90 days based on existing data. Epidemic curves, especially prevalence inflection points with the curves, were predicted successfully for 90 days beyond the data used. Critical to the analysis is the use of EHR data, especially emergency department visits for COVID patients, which is not available with standard public health reporting databases. He will be describing the importance of synthetic data for this analysis, since development of predictive models across geographic regions requires broader data sharing among institutions.

Noa Zamstein, PhD. Dr. Zamstein holds a PhD in computational chemistry and is currently the Senior Data Scientist at MDClone. She is engaged in many of the company’s scientific activities in Israel and North America, with special focus on MDClone’s synthetic solution and its implementation across various sites with the goal of liberating healthcare data. In the past year she has been extensively involved in the NIH-MDClone partnership, supporting various research projects conducted in response to the COVID pandemic. Dr. Zamstein will present the results of analyses characterizing hospitalized COVID patients, and using data available early in admission to predict hospital severity of disease.

Adam Wilcox, PhD, moderator. Dr. Wilcox is currently Chief Analytics Officer at UW Medicine and Professor of Biomedical Informatics and Medical Education at the University of Washington School of Medicine. His research in informatics has focused on applications of informatics principles to improve care, from developing decision support applications, to designing architectural strategies for health information exchange, to creating population health research databases to study issues of health disparities, to methods of improving the use of data and analytics in healthcare. Early in the COVID-19 pandemic, Dr. Wilcox advocated for improved approaches to data sharing to allow shared analytics of emerging datasets, as well as the development of shareable data. He has participated with the group in the analysis of synthetic data and represented the practical need for synthetic data in supporting practical discovery in the COVID pandemic. As moderator, he will focus the discussion of the findings to their broader, combined implications about the utility of synthetic data to support data sharing and broader analytics.

Discussion Questions

- What types of analyses can be done with synthetic data?
- Based on these studies, how do these types of analyses compare with those done with actual patient data?
- What types of analyses can be done with synthetic data but cannot be done with standard de-identified data?
- What are areas where synthetic data could be used to further analyses with healthcare data?
- What are barriers to data sharing that can be addressed by synthetic data? What barriers still remain?

All participants have agreed to participate on the panel at the AMIA 2021 Annual Meeting.
The Women in AMIA-led Podcast: 
Rebranding and Social Media Optimization to Focus on Diversity

Karmen S. Williams, DrPH, MBA1; Mindy K. Ross, MD, MBA2;  
Adela Grando, PhD, MCS3; Tiffany Harman, BSN, RN4; Rachael Howe, MS, RN4;  
Kelly Taylor5; Wendy M. Ingram, PhD6; Davina Zamanzadeh2; Leyla B. Warsame, MD6;  
Zubin Khan, MS7; Anita Murcko, MD3; Omolola Ogunyemi, PhD8;  
1City University of New York, New York, NY, USA; 2University of California Los Angeles,  
Los Angeles, CA, USA; 3Arizona State University, Tempe, AZ, USA; 43M Health  
Information Systems, Inc., Murray, UT, USA; 5American Medical Informatics Association,  
Rockville, MD, USA; 6Geisinger Health, Baltimore, MD, USA; 7The Centers for Disease  
Control and Prevention (CDC), Atlanta, GA, USA; 8Charles R. Drew University of  
Medicine and Science, Los Angeles, CA, USA

Abstract

Women and underrepresented minorities are entering biomedical informatics or achieving leadership roles at a lower rate than other groups. As part of the efforts to address this discrepancy, the Women in AMIA (WIA) Initiative developed the Women in AMIA Podcast in 2017 to showcase women at all stages of their informatics careers. The intent was to increase awareness of women’s contributions to informatics, counterbalance stereotypes, and build the pipeline to leadership. Although the podcast has some reach (an average of 220 downloads per podcast), the podcast leaders wanted to reach to a wider and more diverse audience. To this end, the podcast team partnered with WIA Leadership Program members to rebrand the podcast and launch on a broader scale. In this panel, we will report on the process and results of these efforts including 1) podcast rebranding, 2) social media approach to reach target audience, and 3) podcast and social media metrics of success.

Panel Description

Learning objectives
Understand techniques to: (1) increase podcast audience through branding methods. (2) reach women and underrepresented populations through social media dissemination. and (3) assess podcast impact using social media metrics.

Women and underrepresented minorities enter the clinical informatics pipeline at lower rates and have fewer roles in leadership

Clinical informatics fellowship applicants are majority male and do not include traditionally under-represented groups such as African Americans or those of Hispanic/Latin descent. Those that do enter the healthcare and clinical informatics space are frequently systemically undervalued (i.e., paid less than their counterparts). Women are increasing in numbers in the healthcare space in many domains (e.g., in 2019, 50.5% of medical students are women) but remain underrepresented in healthcare leadership roles, as they are estimated to make up 13% of CEOs and 30% of C-suite roles.

These conditions contribute to women and underrepresented minorities (especially those with intersectionality) not achieving their full potential in the workplace and possibly leaving the field. This can negatively impact business because it has been documented that companies in the top 25 percentile for management diversity were 15-35% more likely to achieve financial returns above their industry mean. Creating a diverse workforce will provide the knowledge and input to address industry pain points.

Podcasting as an educational tool
The digital medium of podcasts has exploded in popularity over the past 15 years. In 2019, at the time of the study, it was estimated that 70 million people had listened to a podcast in the previous month. Approximately 51% of podcast
listeners are male and 63% are white.7 Podcasts have expanded beyond the pop culture realm and into the educational domain.8 The market is expected to grow and expand in upcoming years.

The underrepresentation of women and minorities in clinical informatics is due to many factors, including implicit bias, which leads to stereotyping and results in fewer opportunities.9 Education and role modeling can counteract implicit bias.10 Podcasting can be an effective medium to educate the majority about the talent of underrepresented groups and subvert stereotypes. Podcasts can also foster a connection among underrepresented groups and increase awareness of those with similar or complementary goals.

The Podcast, hosted by Women in AMIA
The WIA Steering Committee started the WIA Podcast in 2017 with a plan to showcase talented women in informatics. As a part of the Pipeline Subcommittee within WIA, the podcast is a part of the larger goal of WIA to attract new women into the field, provide meaningful resources, and to retain women at various career stages. The podcast’s aim was to increase awareness, connect individuals with similar interests, disseminate practical information about how to achieve career success and manage work-life balance.11 The podcasts have featured over 15 women and men at different career stages in informatics and were downloaded close to 3,000 times (on average 220 downloads per podcast). In 2020, the podcast leadership wanted to include all underrepresented groups and broaden the reach of the podcasts and partnered with the WIA Leadership Program Seed Grant Awardees for this endeavor. This effort has included rebranding the podcast (title, logo, music, introduction) and increasing social media efforts. We will report on this process in more detail, including: 1) podcast rebranding, 2) social media approaches to reach women and underrepresented populations as well as increase listenership overall and 3) podcast and social media metrics of success.

Descriptions of Panelists and Presentations
All participants have agreed to take part on the panel.

Karmen S. Williams is an Adjunct Assistant Professor of the City University of New York. Her research focus is in public and population health informatics and increasing representation in STEM fields. She is the director of the podcast, an active member on the WIA Steering Committee, co-chair of the WIA Pipeline Subcommittee, and chair of the Dental Informatics Working Group. As the session moderator, she will provide a brief overview of the WIA led Podcast.

Tiffany Harman is a Nurse Informaticist and Business Development Manager at 3M Health Information Systems. Her focus is clinical terminologies and educating clients on the importance of using standardized data for interoperability and analytics. She was a participant in the WIA Leadership Program and recipient of their Seed Grant award. As a panelist, she will provide a history of the WIA Podcast and the WIA Leadership Program Seed Grant.

Mindy K. Ross is an Assistant Professor at the University of California, Los Angeles David Geffen School of Medicine, Department of Pediatrics. She is a pediatric pulmonologist and physician informaticist. Her research interest is focused on clinical decision support for more personalized asthma management through the patient portal and electronic health record. She was a participant in the WIA Leadership Program and recipient of their Seed Grant award. As a panelist, she will review the rebranding efforts of the podcast.

Maria Adela Grando is an Associate Professor of biomedical informatics at the Arizona State University and Adjunct Assistant Clinical Professor of Medicine at the Mayo Clinic. Her area of expertise in patient-centered technology. She was a participant in the WIA Leadership Program. She is a member of the WIA Podcast Team and a Podcast Host. As a panelist, she will present how to make effective podcasts, resources needed and how to reach your desired audience.

Zubin A. Khan is a Public Health Informatics Fellow at the Centers for Disease Control and Prevention (CDC) and assigned to field location at Utah Department of Health (UDOH). Some of his efforts are focused on the Data Modernization Initiative (DMI) along with Health Interoperability, Population Health and state/federal COVID response. He also serves as a chair-elect at student working group (SWG) at AMIA. As a panelist, he will review the social media marketing strategy and metrics.
Importance of the Panel
It is important to provide outreach to underrepresented women and minorities in clinical informatics to help combat systemic and implicit biases. We hope that disseminating a message of education and awareness will motivate women and underrepresented minorities to engage in leadership roles. By broadening the scope of the WIA-led podcast to include more groups, we aim to educate those in the majority and inspire those in underrepresented groups to choose this field and flourish in it. The WIA-led podcast is an educational resource that will provide the knowledge to influence healthcare leadership to rise into a new era. We hope to educate the audience about the rebranding process and discuss the role that social media plays in highlighting the informatics field. We hope to engage the audience in ideas to increase the reach of the podcast.

Discussion Questions to Enhance Audience Participation
1. Who benefits from the content of the WIA-led Podcast?
2. What could potentially enhance the reach of the WIA-led Podcast?
3. How can we incorporate the WIA-led Podcast in the classroom and broader education?
4. What discussion topics should be added to the WIA-led Podcast?

References
Opportunities to Advance Workflow Automation Using Health Information Technology

Teresa Zayas-Cabán, PhD¹, Grace Cordovano, PhD², J. Marc Overhage, MD, PhD³, Walter Suarez, MD⁴, Tracy Okubo⁵

¹National Library of Medicine, National Institutes of Health, Bethesda, MD; ²Enlightening Results, LLC, West Caldwell, NJ; ³Anthem, Inc, Indianapolis, IN; ⁴Kaiser Permanente, Washington, DC; ⁵Office of the National Coordinator for Health Information Technology, Washington, DC

Abstract

Inefficient health care workflows contribute to the variable quality and poor outcomes sometimes observed in U.S. health care. Increased adoption of health information technology (IT) systems provides opportunities to improve health care workflows through automation. Moreover, broader technology trends offer new possibilities to support more efficient and comprehensive workflows in care delivery. Other sectors making effective use of automation can provide key insights into how to automate health care processes effectively. Capitalizing on this opportunity, the Office of the National Coordinator for Health Information Technology (ONC) has explored opportunities to advance workflow automation in health care through the use of health IT. ONC has focused on the role of organizations, individual stakeholders, and technology to support system-wide transformation that increases efficiency, improves health outcomes, and delivers value. This panel will discuss workflow automation advancements and priorities and related actions needed to propel them further from multiple perspectives including industrial engineering, patients and caregivers, technology development, health care delivery, and health IT policy. Each panelist will summarize their perspective, allowing time for participant interaction. An essential objective of the panel is to discuss opportunities for the informatics community to advance workflow automation priorities and strategies.

Introduction

While the increased adoption and use of health information technology (IT) have created opportunities to more effectively leverage data and knowledge for care delivery, inefficient workflows in health care remain. With continued concerns about the cost, quality, and safety of health care and mounting provider burnout, there is an opportunity to use health IT to enable workflow automation to address these issues.¹ Automation is a means of assisting, augmenting, or replacing human operators with machines or computers to carry out functions that humans customarily perform.² Broader technology trends, such as the rise of the “internet of things” (IoT), advances in artificial intelligence and machine learning, and the evolution of workflow automation tools from assembly belt machinery to a combination of big data, advanced analytics, business intelligence capabilities, improved human-machine interfaces, and digital-to-physical transfer capabilities, have allowed other industries to reduce labor-intensive processes.³⁻⁴ Industries such as agriculture, food production and distribution, manufacturing, transportation, and hospitality have improved quality, productivity, efficiency, and operational safety through workflow automation in recent decades, fueled by modern computational science and democratized access to data.⁵⁻¹¹ Lessons learned from early adopters of automated workflows from these industries may be relevant to health care.

Automating Health Care Workflows through Health IT

The Office of the National Coordinator for Health Information Technology (ONC) has explored opportunities to accelerate workflow automation through the use of health IT.¹² ONC has identified six priorities to advance automation in health care: 1) mobilize nationwide scale automation in near-term “sprints” and long-term “marathons;” 2) enable de-novo discovery of redundant tasks; 3) ensure a ready clinician base for workflow automation; 4) allow all stakeholders to effectively and efficiently engage in health and health care tasks; 5) ensure automation improves patient and caregiver interaction within the health care system; and 6) leverage interoperable health data for automation. ONC has also identified strategies to advance these priorities through specific tactics aimed at educating and convening stakeholders, prioritizing and demonstrating the efficacy of prioritized workflows, and designing incentives and methods to scale successful workflow automation approaches. ONC’s effort has been informed through expert input, a literature review, and a multidisciplinary workshop, emphasizing understanding automation in other industries and identifying applicable lessons learned for health care.¹³ For example, the systems engineering field examines how micro-systems are embedded within macro-systems and operate with varying degrees of complexity.¹⁴
Human factors engineering studies the role of human decision making and action in workflows subject to variation and uncertainty.\textsuperscript{15}

The panel will present: 1) an overview of the factors that drove ONC’s interest in exploring workflow automation; 2) background on workflow automation in health care and other industries; and 3) the perspectives of experts from different fields; and 4) an overview of the identified priorities and actions. The panel members will seek participant feedback on strategies and discuss opportunities to support effective advancement of workflow automation.

**Panel Objectives and Presenters**

This panel aims to share opportunities to advance workflow automation through the use of health IT, and to invite participants to provide reactions and feedback. The panelists will represent a range of perspectives.

Ms. Tracy Okubo (moderator) is a Senior Program Analyst at ONC, where she managed ONC’s workflow automation initiative. She will present the drivers that influenced ONC’s efforts and the resulting priorities. She will also moderate the session.

Dr. Teresa Zayas (panelist) is the Assistant Director for Policy Development at the National Library of Medicine at the National Institutes of Health. While serving as the Chief Scientist at ONC through February 2021, her division led ONC’s scientific efforts including the workflow automation initiative. Through an industrial engineer's lens, she will present the lessons learned and opportunities from non-health care sectors that could be applied to advance health care workflow automation and implementation considerations.

Dr. Grace Cordovano (panelist) is a board-certified patient advocate and the founder of Enlightening Results, LLC. Through her work at Enlightening Results, as well as being the primary caregiver to two disabled adults and a patient herself, she will speak to patients’ and caregivers’ experiences navigating the health care system and opportunities for automation to simplify and improve their interacting with the health care system.

Dr. J. Marc Overhage (panelist) is the Chief Medical Informatics Officer at Anthem who recently led the Population Health Intelligence Strategy at Cerner and served as President and CEO of the Indiana Health Information Exchange. He will discuss how to approach population-level automation opportunities from a health information exchange and care delivery perspective. He will also contribute insights on designing a regulatory approach that advances innovation and observes multiple stakeholders' interests, including clinicians, patients, and payors.

Dr. Walter Suarez (panelist) is the Executive Director of Health IT Strategy and Policy at Kaiser Permanente. He will describe the experience of automating clinical and administrative workflows in an integrated delivery network and offer insights for automation innovation that apply to other delivery models. He will speak to the elements of public policy and market drivers to effectively advance workflow automation.

**Panel Discussion Questions**

The moderator will address the following types of questions to the panelists and encourage the audience to share their input and experience.

1. How can health care apply automation lessons learned from other industries? What factors make health care similar versus unique, and what are the implications for automation?

2. Considering the range of full to partial automation, and the multiple areas in health care where workflow automation can be applied (clinical, administrative, quality, safety), which workflows are suitable to full automation, and where should automation in health care consider retaining a human “in the loop”?

3. Workflow automation holds many promises to improve efficiency, lower costs, and increase safety. How might stakeholder goals, interests and priorities vary, including for patients and caregivers, and what are the implications for implementing the identified priorities?

4. What criteria should the health care sector apply in prioritizing health care workflows to automate?

5. What barriers exist to workflow automation, and how can the priorities address them?
Panel Learning Objectives

1. Participants will learn about opportunities to advance workflow automation.
2. Participants will understand examples of workflow automation in other industries and their potential application to health care.
3. Participants will learn the task and technology characteristics of workflows that make them well-suited for automation across industries.
4. Participants will learn approaches used to design, implement, and evaluate workflow automation.

Conclusion

With the increased availability and exchange of electronic health information through use of health IT, there are numerous opportunities to leverage health data to reduce redundant or inefficient processes via automation. However, additional solutions, resources, and governance may be needed to promote and foster use of health IT to advance automation. By sharing ONC’s efforts to understand opportunities to advance workflow automation, this panel aims to stimulate a rich discussion and gather participant feedback on pursuing impactful automation for multiple stakeholders.

Statement of Participation

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

References

Advancing the Use of FHIR in Research: An Update on NIH’s Efforts

Teresa Zayas-Cabán, PhD1, Belinda Seto, PhD2, Paul Harris, PhD3, Allison Heath, PhD4, Viet Nguyen, MD5

1National Library of Medicine, National Institutes of Health, Bethesda, MD; 2Office of Data Science Strategy, National Institutes of Health, Bethesda, MD; 3Vanderbilt University, Nashville, TN; 4Children’s Hospital of Philadelphia, Philadelphia, PA; 5Stratametrics, Salt Lake City, UT

Abstract

On July 30, 2019, the National Institutes of Health (NIH) issued a notice “to encourage NIH researchers to explore the use of the Fast Healthcare Interoperability Resources® (FHIR®) standard to capture, integrate, and exchange clinical data for research purposes and to enhance capabilities to share research data.” Recently, the Office of Data Science Strategy, NIH, in collaboration with the National Library of Medicine, commissioned an assessment of the landscape of FHIR use in research. The summary findings suggest that FHIR is not yet broadly used for research purposes. There are several challenges from the institutional level to the individual level. Institutions’ decision to implement FHIR involves financial commitment not only to the technology but also to an ongoing need to train their staff. Few researchers currently know how to use FHIR or leverage it in their research studies. This panel will provide an update on NIH’s activities to advance the use of FHIR in research. Panelists will discuss the challenges and opportunities in using FHIR for research with attendees and solicit feedback on NIH’s work to date and input on strategies NIH can implement to better leverage FHIR in research.

Introduction

The increased adoption of health information technology (IT) has increased the availability of electronic health data for clinical care and research. Recent regulations require the use of the Health Level Seven International® (HL7®) Fast Healthcare Interoperability Resources® (FHIR®) standard to enable exchange of clinical and claim information. FHIR provides a standardized way of transmitting health data from one health information system to another through an application programming interface. Use of a standard such as FHIR could accelerate the use of clinical data for research. Additionally, FHIR provides a way to structure data generated from research in a manner that fosters interoperability and interchange of both research and clinical data.

These advancements have created new opportunities for biomedical researchers who may be able to leverage the FHIR standard to both use electronic health data in research as well as make data from a single research study useful for other research endeavors. Data standards are key to the interoperability and reusability of data, allowing data to be more easily analyzed, shared, and combined with other data. Use of data standards in research will help the National Institutes of Health (NIH) achieve its goals in data science and open science.

Advancing the Use of FHIR in Research

In 2019, NIH published a Guide Notice encouraging NIH-funded investigators to explore the use of FHIR to capture, integrate, and exchange clinical data for research purposes and to enhance capabilities to share research data. Subsequently, NIH awarded two related contracts designed to increase the availability of high-quality data using FHIR and published a request for information (RFI) regarding the application of FHIR to research data, including anticipated challenges and opportunities, current experiences, and needed tools. To address feedback received in response to the RFI, NIH is leading an assessment of the current landscape of adoption, implementation, and use of the FHIR standard for research. The landscape assessment includes a mix of targeted expert interviews, federal and stakeholder workshops, and a literature review.

The assessment has found that FHIR is not yet broadly used for research purposes and few researchers currently know how to use FHIR or leverage it in their research studies. The assessment has also uncovered challenges to, and opportunities for, leveraging FHIR in research. Identified challenges to using FHIR in research include: clinical and research needs, technical needs, and legal and regulatory barriers to data access. Some opportunities to leverage FHIR in research include: combining the use of FHIR and artificial intelligence to analyze diagnostic images, use of...
real-world data from electronic health record systems to recruit volunteers for clinical trials, and the ability to interoperate based on standards.

The panel will present an overview of NIH activities to advance FHIR in research to date as well as findings identified from the assessment. The panel members will seek feedback on identified challenges and opportunities and input into strategies NIH can employ to overcome these challenges to formulate its FHIR strategy.

Panel Objectives and Presenters

This panel aims to provide an update on NIH’s activities to advance the use of FHIR in research to date and to invite participants to provide reactions and feedback. The panelists will represent a range of perspectives.

Dr. Teresa Zayas Cabán (moderator and panelist) is the Assistant Director for Policy Development at the National Library of Medicine at the NIH. While on a previous detail at NIH, Dr. Zayas Cabán led the development of NIH’s strategy for the use of FHIR in research. Dr. Zayas Cabán will provide background on NIH’s activities to advance the use of FHIR and other standards in research. She will also moderate the session.

Dr. Belinda Seto (panelist) is the Deputy Director of the Office of Data Science Strategy at NIH, where she also co-leads the NIH FHIR Working Group. Dr. Seto will present on findings from the NIH landscape assessment.

Dr. Paul Harris (panelist) is professor of biomedical informatics and biomedical engineering with extensive experience working in the field of clinical and translational research informatics. He serves as director of the Vanderbilt University Medical Center Office of Research Informatics. Dr. Harris devised and created REDCap, a data collection platform that has seen widespread adoption by more than 4900 institutional partners and over 1.6 million end-users across 141 countries. Dr. Harris will describe methods developed at Vanderbilt and now deployed across the REDCap Consortium allowing research teams to set up automated FHIR-based data exchange between local EHR systems and REDCap. He will also describe current development work and use cases leveraging FHIR for other research-related scenarios, including survey metadata representation, documentation of research consents, and trial data exchange.

Dr. Allison Heath (panelist) is the Director of Data Technology and Innovation at the Center for Data Driven Discovery in Biomedicine at the Children’s Hospital of Philadelphia Research Institute. Dr. Heath has been working to solve big data challenges in the unique context of pediatric diseases and leads the development of the Kids First Data Resource Center. Dr. Heath will present on work she is leading to integrate FHIR into the Kids First Data Resource Center and efforts as part of the NIH Cloud-Based Platform Interoperability (NCPI) FHIR working group. These efforts support clinical and genetic data sharing to in order to accelerate discovery and improve clinical care.

Dr. Viet Nguyen (panelist) is Founder of Stratametrics, board member of HL7, and technical director of the HL7 Da Vinci Project. Dr. Nguyen will provide an understanding of the adoption of FHIR in data exchange between payers, providers and patients and how the adoption of FHIR establishes a foundation for the research community to leverage FHIR in clinical research.

Panel Discussion Questions

The moderator will address the following types of questions to the panelists and encourage the audience to share their input and experience.

1. What are some approaches we can use to overcome challenges to using FHIR in research?
2. What should NIH prioritize in establishing and advancing its FHIR strategy?
3. What are some key opportunities for researchers in using FHIR for their projects?
4. How can the FHIR specification be further developed to better support research needs?
Panel Learning Objectives

1. Participants will learn about NIH’s efforts to advance the use of FHIR in research.
2. Participants will learn about current limitations of the application of the FHIR specification in research.
3. Participants will learn about challenges and opportunities for using FHIR in research.

Conclusion

With the increased availability and exchange of electronic health information through use of health IT, there are numerous opportunities to leverage health data in research by using standards such as FHIR. By sharing NIH’s efforts to advance the use of FHIR in research and findings to date, this panel aims to stimulate a rich discussion and gather participant feedback that will inform the next steps in advancing NIH’s efforts.

Statement of Participation

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

References

A Comparison of Exhaustive and Non-lattice-based Methods for Auditing Hierarchical Relations in Gene Ontology

Rashmie Abeysinghe, PhD\textsuperscript{1}, Fengbo Zheng, PhD\textsuperscript{2}, Licong Cui, PhD\textsuperscript{2,*}
\textsuperscript{1}Department of Neurology, University of Texas Health Science Center at Houston, Houston, TX
\textsuperscript{2}School of Biomedical Informatics, University of Texas Health Science Center at Houston, Houston, TX

Abstract
Uncovering and fixing errors in biomedical terminologies is essential so that they provide accurate knowledge to downstream applications that rely on them. Non-lattice-based methods have been applied to identify various kinds of inconsistencies in different biomedical terminologies. In previous work, we have introduced two inference-based approaches that were applied in an exhaustive manner to audit hierarchical relations in the Gene Ontology: (1) Lexical-based inference framework, and (2) Subsumption-based sub-term inference framework. However, it is unclear how effective these exhaustive approaches perform compared with their corresponding non-lattice-based approaches. Therefore, in this paper, we implement the non-lattice versions of these two exhaustive approaches, and perform a comprehensive comparison between non-lattice-based and exhaustive approaches to audit the Gene Ontology. The domain expert evaluations performed for the two exhaustive approaches are leveraged to evaluate the non-lattice versions. The results indicate that the non-lattice versions have increased precision than their exhaustive counterparts even though they do not capture some of the potential inconsistencies that the exhaustive approaches identify.

1 Introduction
Biomedical terminologies like Gene Ontology, SNOMED CT and NCI thesaurus have received an increase use in terms of knowledge management; data integration, exchange and semantic interoperability; and decision support and reasoning in biomedicine\textsuperscript{1–4}. Biomedical terminologies are constantly curated to reflect the state-of-the-art knowledge of the particular domain that they represent. Though great care is taken to make sure the terminologies reflect the biomedical knowledge accurately, it is inevitable that errors will be introduced due to the manual effort involved in maintaining them combined with how complex terminologies have become over time. It can be laborious to manually audit a modern biomedical terminology due to their size and complexity and hence, automated approaches are preferred.

Non-lattice-based auditing methods have been effectively employed for quality assurance purposes on various biomedical terminologies\textsuperscript{5–11}. These methods focus on subgraph fragments in terminologies that are error prone. In contrast, exhaustive methods focus on the entire terminology without restricting to such subgraphs. Leveraging the lexical features of concepts in non-lattice-subgraphs, various inconsistencies have been uncovered. The advantage of non-lattice-based methods over most other terminology auditing methods is that they are capable of not only identifying errors, but also suggesting remediation measures. Therefore, they require much less manual review effort from domain experts. According to the results of non-lattice-based methods, it is unquestionable that they are effective in uncovering errors in biomedical terminologies. However, a comprehensive comparison of non-lattice-based methods against exhaustive methods has not yet been performed to prove their effectiveness in uncovering inconsistencies.

In previous work, we have developed two exhaustive approaches to audit the hierarchical is-a relations in Gene Ontology (GO): one is a lexical-based inference approach\textsuperscript{12,13}, and the other is a subsumption-based sub-term inference framework\textsuperscript{14}. In this paper, we implement these two approaches on non-lattice subgraphs to detect potential inconsistencies in GO. Leveraging the previous evaluations performed by domain experts, we perform a comparison of the effectiveness of the exhaustive version and non-lattice version for both approaches.

*Corresponding author. Email: licong.cui@uth.tmc.edu
2 Background

2.1 Methods to audit biomedical terminologies

Many approaches have been proposed to audit biomedical terminologies\textsuperscript{15}. Abstraction networks which are summary graphs of terminologies have been extensively explored to perform quality assurance\textsuperscript{16–18}. Zhe et al. have worked on identifying trapezoid structures in the hierarchies of a pair of terminologies to identify missing concepts\textsuperscript{19}. Bodenreider has introduced a method to uncover missing \textit{is-a} relations through inference of logical definitions constructed using lexical features of concept names\textsuperscript{20}. Zheng et al. have introduced a transformation-based method that leverages Unified Medical Language System (UMLS) knowledge to identify missing hierarchical relations in terminologies in the UMLS\textsuperscript{21}. Peng et al. have proposed a new algorithm to predict new GO terms and connect them to existing GO\textsuperscript{22}. Mougin et al. have reasoned over relationships to identify redundant relations in GO and detected missing relations by using compositional structure of the concept names\textsuperscript{23}. Xing et al.’s work combined dynamic programming with topological sort to detect redundant relation in biomedical terminologies including GO\textsuperscript{24}. More recently, deep learning has been explored to audit biomedical terminologies. Zheng et al. have proposed a method that leverage deep learning to predict the concept names of new concepts that comply with the naming convention of the terminology\textsuperscript{25}. Liu et al. have introduced a deep learning approach that can predict the placement of a new concept in the hierarchy of SNOMED CT\textsuperscript{26}.

2.2 Non-lattice subgraphs

Being a lattice is considered a desirable property for a well-formed terminology\textsuperscript{5,27,28}. A terminology forms a lattice if any pair of concepts have a unique maximal shared descendant and a unique minimal shared ancestor. A pair of concepts is known as a non-lattice pair, if they have more than one maximal shared descendant or minimal shared ancestor. For example, in Figure 1, concepts $A$ and $B$ form a non-lattice pair since they share two maximal common ancestors $E$ and $F$.

The non-lattice pair $(A, B)$ defines a non-lattice subgraph as follows. First, the maximal common descendants ($mcd$) of the non-lattice pair is obtained. For the non-lattice pair $(A, B)$, the maximal common descendants $mcd(A, B) = \{E, F\}$. Then we reversely compute the minimal common ancestors ($mca$) of concepts $E$ and $F$. This yields us $mca(mcd(A, B)) = \{A, B, C\}$. Then all the concepts and relations between $mcd(A, B)$ and $mca(mcd(A, B))$ is aggregated to form the non-lattice subgraph. This yields a non-lattice subgraph with six concepts $\{A, B, C, D, E, F\}$.

![Figure 1: An example non-lattice subgraph.](image)

2.3 Non-lattice-based auditing methods

Previously, we have investigated a number of non-lattice-based methods to audit biomedical terminologies. Four lexical patterns in non-lattice subgraphs were investigated to uncover missing \textit{is-a} relations and missing concepts in SNOMED CT\textsuperscript{5}. The same approach was applied to NCI thesaurus introducing two more lexical patterns\textsuperscript{6}. We also introduced a method that leverages enriched lexical attributes of concepts in non-lattice subgraphs in SNOMED CT to uncover \textit{is-a} relation inconsistencies\textsuperscript{7}. We applied similar approaches on the NCI thesaurus while also investigating the inheritance of lexical attributes from all ancestors and leveraging role definitions of concepts.

Next we discuss the two exhaustive methods from previous work that we will compare later with their non-lattice-based counterparts.
2.4 Lexical-based inference framework

To audit *is-a* relations in GO, we introduced a lexical-based inference framework. In this approach, we represented the name of each concept using two models: set-of-words and sequence-of-words. For each model, we generated hierarchically-linked and -unlinked Partial Matching Concept-Pairs (PMCPs), \((A, B)\), such that both \(A\) and \(B\) have the same number of words, and contain at least one word in common and a fixed number of different words \((n = 1, 2, 3, 4, 5)\). The linked and unlinked concept-pairs further infer corresponding linked and unlinked term-pairs (ITPs) respectively. If the same ITP is inferred by a linked and unlinked PMCP, this is considered as an inconsistency. Applying this approach to the March 28, 2017 release of GO, a total of 5,359 potential inconsistencies were found by the set-of-words model and 4,959 were found by the sequence-of-words model. A random sample of 250 potential inconsistencies identified through this method was evaluated by domain experts to validate their correctness. The results showed that the set-of-words model achieved a precision of 53.78% while the sequence-of-words model achieved a precision of 57.55%.

2.5 Subsumption-based sub-term inference framework

In previous work, we developed a subsumption-based sub-term inference framework (SSIF) to audit GO. In SSIF, we represented each concept \(A\) with a sequence-based representation \(E(A) = [e_1, e_2, e_3, ..., e_n]\), where each element is either a word or a sub-concept. We leveraged part-of-speech tagging, sub-concept matching and antonym tagging to construct the sequence-based representation for each concept. Then, we introduced three conditional rules: Monotonicity, Intersection, and Sub-concept that utilized the sequence-based representation to uncover problematic *is-a* relations. Here, we briefly discuss the three rules.

Monotonicity rule suggests \(A\ is-a\ B\) if both \(A\) and \(B\) have the same number of elements in their sequence-based representations \(E(A)\) and \(E(B)\) respectively, their corresponding elements \(A_i\) and \(B_i\) are either equal or if they are sub-concepts (discussed below) \(A_i\ is-a\ B_i\), and \(E(A)\) does not contain an element which is an antonym of any element of \(E(B)\).

Intersection rule suggests a missing *is-a* relation between a concept \(A\) and an intersecting concept \(X\) as follows. Suppose that \(A\) has a pair of ancestors \(B\) and \(C\). The intersecting concept \(X\) of \(B\) and \(C\) is defined as another concept that contains the lexical properties of both \(B\) and \(C\) and is lexically the most general concept that is a descendant of both \(B\) and \(C\). Therefore, since \(A\) is also a descendant of both \(B\) and \(C\), this rule suggests that \(A\ is-a\ X\). Note that \(A\) should not have an element which is an antonym of an element of \(X\).

Sub-concept rule suggests a missing *is-a* relation among a concept and its sub-concept. We say that a concept \(B\) is a sub-concept of a concept \(A\) if \(B\) is a proper substring of \(A\). Suppose that if the sub-concept \(B\) is the last element of \(A\), all other elements of \(A\) are either sub-concepts or belong to parts-of-speech noun or adjective, and \(A\) does not have an element which is an antonym of any element of \(B\), then the Sub-concept rule suggests \(A\ is-a\ B\).

Applying SSIF to the October 3, 2018 release of GO, 819; 691; and 669 potential inconsistencies were uncovered for Monotonicity, Intersection, and Sub-concept rules respectively. Domain experts evaluated a random sample of 210 potential inconsistencies uncovered by SSIF and the results showed that SSIF achieved a precision of 60.61%, 60.49%, and 46.03% for Monotonicity, Intersection, and Sub-concept rules respectively.

3 Methods

We first extract all the non-lattice subgraphs from both the March 28, 2017 and October 3, 2018 releases of Gene Ontology (the same versions used in the exhaustive approaches) leveraging an efficient large-scale non-lattice detection algorithm. Then we develop the non-lattice version of the two approaches as follows.

3.1 Non-lattice lexical-based inference framework

For both set-of-words model and sequence-of-words model, the non-lattice version of the lexical-based inference framework is as follows. A pair of hierarchically-linked or -unlinked concepts \((A, B)\) form a non-lattice PMCP if both \(A\) and \(B\) have the same number of words, contain at least one word in common and a fixed number of different
words \( n = 1, 2, 3, 4, 5 \), and are in the same non-lattice subgraph. From a non-lattice PMCP, a non-lattice ITP is derived by removing the common words across the two PMCPs. We say that an inconsistency exists in a non-lattice subgraph if it contains a linked non-lattice PMCP and an unlinked non-lattice PMCP, which infers the same non-lattice ITP. Figure 2 shows an example non-lattice subgraph that exhibits this scenario. Note that the concepts of linked PMCP are in green and the concepts of the unlinked PMCP are in red. This can be obtained by both the set-of-words and sequence-of-words models.

3.2 Non-lattice subsumption-based sub-term framework

Exhaustive SSIF proposed three conditional rules: Monotonicity, Intersection, and Sub-concept to uncover potential is-a inconsistencies. Here, we define their non-lattice-based counterparts as follows.

**Non-lattice-based Monotonicity rule** suggests \( A \) is-a \( B \) if concept \( A \) and concept \( B \) satisfy the following conditions:

- they are in the same non-lattice subgraph;
- they have the same number of elements in their sequence-based representations \( E(A) \) and \( E(B) \) respectively;
- their corresponding elements \( A_i \) and \( B_i \) are either equal or if they are sub-concepts, both of them are in the above-mentioned non-lattice subgraph where \( A_i \) is-a \( B_i \); and
- \( E(A) \) does not contain an element which is an antonym of any element of \( E(B) \).

Figure 2: A missing is-a relation (dashed link in red) identified by the non-lattice lexical-based inference framework. Concept pair of linked PMCP is in green and concept pair of unlinked PMCP is in red. The suggestion here is that peptidyl-threonine trans-autophosphorylation is-a peptidyl-threonine autophosphorylation.

Figure 3: A missing is-a relation (dashed link in red) identified by the Monotonicity rule of the non-lattice SSIF. Concepts in green exists as sub-concepts of concepts in red. The suggestion here is positive regulation of secondary heart field cardioblast proliferation is-a positive regulation of cardioblast proliferation.
Figure 3 shows a non-lattice subgraph exhibiting Monotonicity rule. This example shows a suggested missing is-a relation between concepts $A = \text{positive regulation of secondary heart field cardioblast proliferation}$ and $B = \text{positive regulation of cardioblast proliferation}$. Note that with the sequence-based representation, concept $A$ here is represented as $E(A) = \{\text{positive, (regulation of secondary heart field cardioblast proliferation)}\}$, i.e., the first element is positive and the second element is regulation of secondary heart field cardioblast proliferation which is a sub-concept. Similarly, concept $B$ is represented as $E(A) = \{\text{positive, (regulation of cardioblast proliferation)}\}$. Since $A$ and $B$ are both in the same non-lattice subgraph, have two elements each, their first elements are equal and the second elements are sub-concepts having an is-a relation which also exists in the same non-lattice subgraph (denoted by green circles), we suggest a missing is-a relation among these two concepts (denoted by red dashed line).

Non-lattice-based Intersection rule suggests a missing is-a relation as follows. Suppose that $A$ has a pair of ancestors $B$ and $C$ whose intersecting concept is $X$. If all $A, B, C$, and $X$ are in the same non-lattice subgraph, this rule suggests a missing relation $A$ is-a $X$ if $A$ does not have an element that is an antonym of an element of $X$. Figure 4 shows a non-lattice subgraph exhibiting Intersection rule. For instance, concepts $B = \text{regulation of establishment of cell polarity}$ and $C = \text{regulation of establishment or maintenance of cell polarity regulating cell shape}$ (denoted by green circles) which are in the same non-lattice subgraph has an intersecting concept $X = \text{regulation of establishment of cell polarity regulating cell shape}$ which is also in the same non-lattice subgraph. Concept $A = \text{regulation of establishment of bipolar cell polarity regulating cell shape}$ which also exists in the same non-lattice subgraph is a descendant of both $B$ and $C$. Therefore, we suggest a missing is-a relation between the concepts $X$ and $A$ (denoted by the red dashed link).

Non-lattice-based Sub-concept rule suggests a missing is-a relation between a concept $A$ and its sub-concept $B$ if

- both of $A$ and $B$ are in the same non-lattice subgraph;
- the sub-concept $B$ is the last element of $A$, and all other elements of $A$ are either a sub-concept or noun or adjective; and
- $A$ does not have an element which is an antonym of any element of $B$.

For instance, Figure 5 contains a non-lattice subgraph exhibiting Sub-concept rule. Here, concept $A = \text{positive regulation of phenotypic switching by regulation of transcription from RNA polymerase II promoter}$ is represented by the sequence-based representation as $E(A) = \text{positive, (regulation of phenotypic switching by regulation of transcription from RNA polymerase II promoter)}$. i.e. the last element of $A$ is a subconcept $B = \text{regulation of phenotypic switching by regulation of transcription from RNA polymerase II promoter}$ which also exists in the same non-lattice subgraph. The remaining element positive is an adjective. Therefore, based on the non-lattice Sub-concept rule, we suggest there exists a missing is-a relation between $A$ and $B$ (denoted by the dashed red line).
3.3 Evaluating non-lattice-based methods

To evaluate the performance of the non-lattice versions of lexical-based inference framework and SSIF, we leverage the evaluations performed on their exhaustive counterparts. If a potential inconsistency identified through the non-lattice approach exists in the evaluation sample of the exhaustive approach, then, that particular sample is considered for evaluating the non-lattice approach. So, we simply take the intersection between the non-lattice results with the exhaustive evaluation set to identify the evaluation samples obtainable by the non-lattice approach. Then, based on the number of valid inconsistencies and false positives, we compute the precision.

4 Results

In this section, we compare the potential inconsistencies uncovered by the exhaustive and non-lattice versions of the lexical-based inference framework and SSIF.

Table 1 shows potential inconsistencies identified by exhaustive and non-lattice lexical-based inference frameworks. For example, the exhaustive approach has identified 5,359 potential inconsistencies with the set-of-words model while the non-lattice-based approach has identified 1,875 potential inconsistencies with the same model.

Table 1: A comparison of all potential inconsistencies uncovered by exhaustive and non-lattice lexical-based inference frameworks.

<table>
<thead>
<tr>
<th></th>
<th>Set-of-words</th>
<th>Sequence-of-words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exhaustive</td>
<td>5,359</td>
<td>4,959</td>
</tr>
<tr>
<td>Non-lattice</td>
<td>1,875</td>
<td>1,691</td>
</tr>
</tbody>
</table>

Table 2 displays potential inconsistencies uncovered by the three rules of exhaustive and non-lattice versions of SSIF. For instance, the exhaustive version of Monotonicity rule identified 819 potential inconsistencies while the non-lattice version of the Monotonicity rule uncovered 354 potential inconsistencies.

Table 2: A comparison of all potential inconsistencies uncovered by exhaustive and non-lattice SSIF.

<table>
<thead>
<tr>
<th></th>
<th>Monotonicity</th>
<th>Intersection</th>
<th>Sub-concept</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exhaustive</td>
<td>819</td>
<td>691</td>
<td>669</td>
</tr>
<tr>
<td>Non-lattice</td>
<td>354</td>
<td>679</td>
<td>75</td>
</tr>
</tbody>
</table>

Table 3 shows a comparison of performance between the exhaustive and non-lattice versions of lexical-based inference.
framework. For example, the exhaustive version achieved a precision of 53.78% when the set-of-words model was used. The non-lattice-based approach achieved a precision of 58.97% for the set-of-words model.

| Table 3: A comparison of the performance of exhaustive and non-lattice lexical-based inference framework. |
|---------------------------------------------------------|---------------------------------------------------------|---------------------------------------------------------|
| **No. of potential inconsistencies** | **No. of valid inconsistencies** | **Precision** |
| **Exhaustive** | **Non-lattice** | **Exhaustive** | **Non-lattice** | **Exhaustive** | **Non-lattice** |
| Set-of-words | 238 | 39 | 128 | 23 | 53.78% | 58.97% |
| Sequence-of-words | 212 | 28 | 122 | 20 | 57.55% | 71.43% |

Table 4 shows a performance comparison of the exhaustive and non-lattice versions of SSIF. For instance, the exhaustive approach achieved a precision of 60.61% with the Monotonicity rule while the non-lattice-based approach achieved 61.54%.

**Table 4: A comparison of the performance of exhaustive and non-lattice SSIF.**

<table>
<thead>
<tr>
<th><strong>No. of potential inconsistencies</strong></th>
<th><strong>No. of valid inconsistencies</strong></th>
<th><strong>Precision</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Exhaustive</strong></td>
<td><strong>Non-lattice</strong></td>
<td><strong>Exhaustive</strong></td>
</tr>
<tr>
<td>Monotonicity</td>
<td>99</td>
<td>39</td>
</tr>
<tr>
<td>Intersection</td>
<td>81</td>
<td>79</td>
</tr>
<tr>
<td>Sub-concept</td>
<td>63</td>
<td>8</td>
</tr>
</tbody>
</table>

5 Discussion

In this paper, we implemented the non-lattice versions of our previous exhaustive lexical-based inference framework and subsumption-based sub-term inference framework for auditing *is-a* relations in GO, and performed a comparison between exhaustive and non-lattice-based approaches.

5.1 Performance comparison

From the performance comparison in Table 3 for the lexical-based inference approach, it can be seen that the non-lattice approach has outperformed the exhaustive approach considerably. The precision of the non-lattice-based approach is better by 5.19% for set-of-words model and 13.88% for the sequence-of-words model. However, the non-lattice-based approach only uncovers 35% and 34% of the potential inconsistencies that the set-of-words and sequence-of-words of the exhaustive approach identifies respectively. For instance, by both the set-of-words and sequence-of-words models, the exhaustive version identifies the missing *is-a* relation *diadenosine polyphosphate catabolic process* *is-a* *small molecule catabolic process* which the non-lattice-based approach does not.

From Table 4, it can be seen that the non-lattice-based approach exceeds the precisions by 0.93% for Monotonicity, 1.54% for Intersection and 16.47% for Sub-concept rules. However, it can also be seen that the non-lattice-based approach identifies 42%, 96%, and 11% of the potential inconsistencies that Monotonicity, Intersection, Sub-concept rules of the exhaustive approach finds respectively. For example, with the Sub-concept rule, the exhaustive version identifies the missing *is-a* relation *skeletal muscle cell differentiation* *is-a* *muscle cell differentiation* that the non-lattice version does not.

Therefore, from the results it is clear that the two non-lattice-based approaches perform better than their exhaustive counterparts in terms of precision. It seems that this performance increase also depends on the method itself since some methods (e.g. Sub-concept rule of SSIF) has gained more than the others.

It can be also seen that non-lattice-based approaches miss some of the potential inconsistencies that exhaustive approaches are able to uncover. We did not expect non-lattice subgraphs would capture all kinds of inconsistencies that exists in biomedical terminologies. However, we expected that the concentration of errors in non-lattice subgraphs would be higher than that in the general terminology and therefore, if the same method is applied exhaustively and inside non-lattice subgraphs, the non-lattice version would achieve a better precision.
It is easier to analyze a non-lattice subgraph than analyzing the entire terminology. If non-lattice subgraphs have a higher concentration of errors, such analysis may lead to the identification of new types of inconsistencies. Strategies to uncover and fix inconsistency types identified by analyzing non-lattice subgraphs may be applied exhaustively to identify more inconsistencies that non-lattice subgraphs itself do not capture. Therefore, non-lattice-based approaches may influence future exhaustive approaches in turn uncovering much more inconsistencies than NLS approaches alone can uncover.

5.2 Level differences of concepts and their superconcepts in the potential missing is-a identified

Since the non-lattice-based methods focus on substructures of a terminology and exhaustive methods works on the entire terminology, we also performed a level-based analysis of the potential missing is-a relations identified by exhaustive and non-lattice-based methods. For a descendant and an ancestor of a potential missing is-a relation, we first computed the level in the hierarchy for the descendant and the ancestor. The level of a particular concept is the number of concepts in the longest path from root to the particular concept. For a particular potential missing is-a relation, we subtract the level of the ancestor from the descendant to get the level difference.

Figure 6 shows a plot of the level differences and the number of is-a relations for the lexical-based inference approach. It can be seen that the lexical-based inference approach favors low level differences for both exhaustive and non-lattice-based methods. This means that the concepts in the potential missing is-a relations tend to be closer to each other in the hierarchy of the ontology. Other than the set-of-words model in the non-lattice-based approach, all other models have most number of potential missing is-a relations with a level difference of 0.

![Figure 6: Distribution of level differences of descendant and ancestor in potential inconsistencies uncovered through exhaustive and non-lattice version of lexical-based inference framework.](image)

Figure 7 displays the plot for level differences and the number of is-a relations for SSIF. It seems SSIF also tends to suggest hierarchically closer concepts as potentially missing is-a. Other than the Sub-concept rule in the exhaustive approach, all others have most number of potentially missing is-a with a level difference of 1.

5.3 Future work

In this work, we leveraged two previous exhaustive methods to compare non-lattice-based and exhaustive methods. We leveraged evaluations from the exhaustive methods to evaluate the non-lattice-based methods. Because of this, the
In this work, we also focused on one terminology: Gene Ontology. In the future we expect to see whether these results are terminology specific by performing a similar comparison on other major biomedical terminologies like SNOMED CT and NCI thesaurus.

6 Conclusion

In this paper, we performed a comparison of exhaustive methods vs non-lattice-based methods for auditing biomedical terminologies. We implemented non-lattice versions of two of our previous exhaustive works in auditing the Gene Ontology: (1) lexical-based inference framework and (2) subsumption-based sub-term inference framework. We leveraged the domain expert evaluations performed for the exhaustive methods to evaluate the non-lattice-based methods. The results indicate that non-lattice-based methods achieved better precisions than their exhaustive counterparts though, they do not capture some of the potential inconsistencies identified by exhaustive approaches.

Acknowledgment

This work was supported by the National Science Foundation (NSF) through grant IIS-1931134, and the National Institutes of Health (NIH) through grants R01LM013335 and R01NS116287. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NSF or NIH.

References

4. Bodenreider O. Biomedical ontologies in action: role in knowledge management, data integration and decision support.


Detecting Fine-Grained Emotions on Social Media during Major Disease Outbreaks: Health and Well-being before and during the COVID-19 Pandemic

Olanrewaju Tahir Aduragba, MSc1, Jialin Yu, MSc1, Alexandra I. Cristea, PhD1 and Lei Shi, PhD1

1Department of Computer Science, Durham University, Durham, United Kingdom

Abstract

The COVID-19 pandemic has affected the whole world in various ways. One type of impact is that communication, work, interaction, a great part of our lives has moved online on various platforms, with some of the most popular being the social media ones. Another, arguably less visible impact, is the emotional impact. Detecting and understanding emotions is important, to better discern the emotional health and well-being of the global population. Thus, in this work, we use a social media platform (Twitter) to analyse emotions in detail. Our contribution is twofold: (1) we propose EmoBERT, a new emotion-based variant of the BERT transformer model, able to learn emotion representations and outperform the state-of-the-art; (2) we provide a fine-grained analysis of the pandemic’s effect in a major location, London, comparing specific emotions (annoyed, anxious, empathetic, sad) before and during the epidemic.

Introduction

Disease outbreaks have remained a problem over many years. Most recently, with the Coronavirus disease (COVID-19) pandemic rapidly evolving, people across the globe have become extremely vulnerable. As of March 2021, over 100 million people from more than 200 countries have been infected1. To suppress the virus, governments worldwide have had to introduce restrictions to human movement and social gatherings. The widespread disruption to human lives has led to stress, economic hardships, and uncertainties about the future2. Therefore, it has stirred a diverse range of emotional responses, such as anxiety, sadness and anger3. In the absence of face-to-face undertakings and meetings, people have resorted to using social media to express how they feel. Social media sites (e.g., Twitter) are essential platforms, where users self-report their thoughts and feelings during public health emergencies4, allowing at the same time access to the public emotional health and well-being pulse. As a result, user-generated contents have become useful to monitor public perceptions and sentiments during past disease outbreaks, such as Ebola and Zika epidemics5,6, and the current COVID-19 pandemic7.

Detecting fine-grained human emotions from text is a challenging research area, due to limited manually annotated data. Therefore, emotional cues, such as emoticons and hashtags from texts, have been used for distant supervision, to serve as emotion labels, to build powerful deep learning models and predict fine-grained emotions accurately. For example, hashtags, such as #sad, #angry and #happy, were used to automatically annotate general Twitter data with fine-grained emotions and further used to train models, to learn useful text representations in an emotional context8.

Although pre-trained language models have been successful in a diverse range of natural language processing tasks9, they fail to consider knowledge that is important for tasks related to the determination of emotions, valence, or affective states from text10. For instance, previous studies11,12 have shown that learning sentiment-specific knowledge during pre-training could enhance the understanding of the sentiments in text, thereby improving sentiment analysis performance. However, there is limited research on the use of emotion-specific knowledge in fine-grained emotion detection. As sentiment analysis is related to emotion detection, we argue that integrating emotional knowledge into pre-trained models will result in better performance for fine-grained emotion detection. To this end, our research questions are:

**RQ1**: Can incorporating emotion information in learning models increase the performance of these models to understand the effects of major health events, such as pandemics?

**RQ2**: How can fine-grained emotion detection related to emotional health and well-being be performed based on social media utterances?
Thus, our work presented in this paper makes the following contributions:

1. We propose a novel deep transfer learning framework modelling entity-invariant emotional representations, and EmoBERT, the model we learnt under this framework. EmoBERT infuses emotion-specific knowledge into a pre-trained language model (BERT), thus being capable of learning useful emotional representations, unlike those obtained from traditional pre-trained language models. We experiment and compare against several state-of-the-art approaches, to illustrate the importance of integrating such knowledge in pre-trained language models for predicting fine-grained emotions (i.e. detecting a specific emotion, e.g. 'sad', rather than its presence).

2. We conduct, to the best of our knowledge, the first study on how the COVID-19 pandemic has affected public emotions on Twitter users in London, United Kingdom. We compare annoyed, anxious, empathetic and sad emotions expressed in tweets from March 2020 with the same period in 2019. We also separately perform an analysis of the hashtags mostly used in tweets expressing the selected emotions.

The remainder of the paper is organised as follows: the Related Work section gives an overview of related literature in emotion detection. The Model section details our proposed approach. The Results section reports experimental results and findings. The Conclusion section presents the conclusions and directions for future research.

Related Work

Emotion detection has been extensively studied in various settings, including online health-related forums and social media sites. In particular, analysis of emotions on social media sites has attracted significant research in the public health domain, to monitor public sentiments during disease outbreaks. However, emotion detection during public health emergencies is challenging, due to the lack of annotated data. Previous studies focused on using existing resources that were developed for general domains. Due to the nature of social media texts, such analyses could be biased towards domain-general contexts. To address the problem with limited availability of labelled data in emotion detection, some studies experimented with automatically annotating user-generated data with cues (e.g., emojis, emoticons, hashtags) specific to such corpora. However, there is limited research on the usage of emojis as emotion labels in the context of public health emergencies. Based on this, we have focused on analysing emojis in tweets - as we arguably assume that a tweet expresses an emotion if it contains an emoji.

Furthermore, existing research proved that pre-trained language models achieve state-of-the-art performances on natural language processing tasks, such as text classification, named entity recognition and question answering. Such models’ standard workflow involved initially pre-training on large amounts of unlabelled corpus data using a language model loss function, and the pre-trained model would then be fine-tuned on labelled data, to adapt to an (often different) downstream task. While being successful in the analysis of sentiments and emotions, they failed to consider task-specific objectives that may improve performance. A recent research enriched their model with sentiment information, by masking sentiment words and predicting the masked words and managed thus to capture rich sentiment knowledge. Another related work aimed to incorporate sentiment knowledge, by masking emoticons in a text and predicting if the masked token was an emoticon or not. However, most of these techniques focused mainly on identifying positive or negative sentiments. Our work uses similar concepts to infuse information about a range of emotions, but with the aim to benefit fine-grained emotion detection.

Model

In this section, we describe our fine-grained emotion detection in tweets model: EmoBERT. The success of deep neural networks (DNNs) has provided the ability to learn useful representations from large data (with or without annotation). DNNs have been widely used in natural language processing. The captured knowledge can be leveraged and used in downstream tasks. As a result, pre-trained language models, such as Bidirectional Transformers for Language Understanding (BERT), have been empirically successful across various tasks. BERT utilised Transformer networks and was trained on large amounts of unlabelled data from the Bookcorpus (containing 800 million words) and English Wikipedia (containing 2.5 billion words). Other variants of BERT, such as XLNet and RoBERTa,
have been proposed since the launch of BERT. We thus employ a BERT-based architecture, infused with emotional knowledge, for our task of fine-grained emotion detection (detecting a specific emotion e.g. sad or anxious) in tweets. We aim to instil in our model a strong inductive power and learn useful emotional knowledge. We exploit unlabelled tweets data related to our target domain to learn this knowledge. By adapting the model to the data closer to the distribution of our target data, our model will be capable of learning emotion-related bias. Our approach follows the standard BERT workflow, pre-training and fine-tuning. However, our fine-tuning step involves an intermediate step that helps incorporate emotional knowledge into our model (see Figure 1).

Pre-Training

Pre-trained language models have become popular in recent years, because they became capable of learning knowledge that is useful for transfer learning in natural language processing. One of the best known pre-trained language models, as aforementioned, is BERT. It was pre-trained using two language modelling objectives: (1) masked language modelling (MLM), where randomly masked tokens are predicted, and (2) next sentence prediction (NSP), predicting two input sentences that are next to each other. The pre-trained BERT could then be used to fine-tune on downstream tasks, such as text classification. Due to limited computational power, we employ BERT-base, a version of BERT which contains 12 transformer layers, to initialise EmoBERT. As noted, we expect the general knowledge captured from pre-training BERT to be useful for our fine-tuning step.

Emotion Knowledge Enhanced Fine-tuning

The first fine-tuning stage is the major novelty of our approach. Our goal in this intermediate fine-tuning step is to enhance our model with emotional knowledge for emotion detection. Formally, given our target domain $D_T$, our source domain $D_S$, and the tweet representations $X_S \in D_S$ belonging to the source domain, we aim to learn the target domain tweet representations $X_T$ via modelling their marginal probability distribution $P(X_T)$ over the target domain $D_T$ with explicit modelling of emotion knowledge. Note that both $X_S$ and $X_T$ consist of $N$ tweets and can be denoted in the general form as $X = \{x_1, x_2, \ldots, x_N\}$. At this stage, we fine-tune BERT, by using the mask language modelling objective to recover emotion information, while learning about the tweet representations $X_T$ and the distribution of our domain data $P(X_T)$. Our intermediate fine-tuning task is based on two concepts: (1) extraction of tweets with emotion emojis, and (2) emotion word masking.

Figure 1: EmoBERT Architecture
Extraction of Tweets with Emotion Emojis

As discussed, we assume that if a tweet contains an emotion emoji, then it carries some emotional information. Facial expression emojis allow Twitter users to express their feelings with non-verbal elements in a tweet\(^9\). Although emojis in texts may not always reflect the emotions experienced by users, recent works have shown that they can still be used to classify the emotional content of a text\(^30,34\) accurately. For example, sad emojis such as 😢 and 😞 show a strong correlation with sad emotions in tweets, whilst anger emojis such as 😡 and 😡 are strongly associated with anger emotions\(^31\). Our use of tweets with emotion emojis has two advantages. Firstly, Twitter is very noisy, and it contains many tweets that do not express any emotion. This way, we can eliminate spurious tweets and preserve only tweets with, arguably, emotional information. Secondly, as discussed above, we can take advantage of the emotional cues from the emojis present and use this information to enhance the detection of emotions from tweets in our model. We use emojis that belong to the “Smileys & Emotion” category based on the official Unicode emoji set\(^1\), that are part of the most commonly used emojis on Twitter\(^2\), to filter out our unlabelled tweets datasets. Thus, we implicitly induce emotion specific biases, by only training on tweets that contain at least one emoji from our emoji shortlist.

Emotion Word Masking

We use the same MLM objective function as in the BERT paper\(^9\) to fine-tune the pre-trained model extracted from the Huggingface library\(^32\). The MLM objective was to predict the original tokens, given the masked token. BERT’s standard masking strategy involved randomly sampling and selecting 15% of the input tokens and then replacing 80% of these sampled tokens with a special masked token [MASK]. Another 10% were replaced with a random token, while the remaining 10% of the sampled tokens were kept unchanged. In addition to the standard MLM for BERT, our approach uses emotion word masking (see Figure 2) for learning emotional knowledge in the text. This masking process is different from the standard MLM in BERT. Since the emotion words are likely to impact on the emotion expressed in a tweet, we use a higher masking probability (50%) for masking emotion words.

To determine emotion words, we employ the NRC Word-Emotion Association Lexicon (a.k.a. EmoLex)\(^33\). EmoLex consists of 14,182 crowdsourced English words associated with eight basic emotions: anger, anticipation, disgust, fear, joy, sadness, surprise and trust\(^34\). Whilst there are other lexical emotion resources, such as DepecheMood\(^35\), we chose EmoLex due to its large size and coverage of broad emotional dimensions\(^36\). Additionally, past works such as\(^37\) have used this lexicon as a prior association of emotions, to detect emotions in texts automatically. An initial examination of the lexicon shows that some lexicon words are associated with no emotion category mentioned above. We remove such words, since they were deemed not useful for our emotion-word masking and select only words associated with at least one emotion on our list. After the elimination process, we use a total of 4,463 words from the lexicon.

![Figure 2: Emotion word masking](https://emojitracker.com)

---

\(^1\)https://unicode.org/emoji/charts/full-emoji-list.html

\(^2\)as indicated in https://emojitracker.com
BERT used the WordPiece algorithm as its tokenisation algorithm, to deal with words that are out of its vocabulary, by splitting them into sub-word tokens present in its vocabulary. We expected that some emotional words would be out of BERT’s vocabulary. For example, the emotion word “somatic” could be split into “so” and “##matic” by this BERT tokeniser. In our model, we define the masked emotion word as the sub-word tokens that correspond to the original emotion word.

Since emotion words appear in emotional contexts, our model aims to capture implicit emotion knowledge representations and preserve information that could be useful in detecting the emotions expressed in a tweet. For the intermediate fine-tuning task, we use a collection of original (no retweets) English tweets collected in April, 2020, related to COVID-19, using keywords such as coronavirus, corona, covid, covid-19, coronaoutbreak, 2019nCoV, pandemic, epidemic, wuhanandlockdown and sars-cov-2. Consequently, we filter the training tweet datasets to contain at least one emoji from our emoji selection and remove all duplicate tweets. In total, our training dataset contains 1,540,983 tweets after the data filtering process. Furthermore, we pre-process all tweets, by replacing all Twitter usernames and URLs with the common tokens: <user> and <url>, respectively.

**Emotion Detection Fine-tuning**

In the final step, we fine-tune EmoBERT on our downstream task: fine-grained emotion detection. Outputs from the emotion knowledge-enhanced fine-tuned model are trained to classify the emotion of a tweet. Following the fine-tuning setting in the original BERT paper, we use the last state vector of the classification token [CLS] as input and feed it into a feed-forward neural network, to predict the respective emotion.

**Experiment**

We use a publicly available annotated dataset, SenWave, providing fine-grained emotion labels of tweets during the COVID-19 pandemic, to evaluate our model. This dataset consists of 10,000 English tweets, labelled with 10 categories: optimistic, thankful, empathetic, pessimistic, anxious, sad, annoyed, denial, official report and joking. We consider 3 emotions (annoyed (anger), anxious and sad), which are part of the fundamental emotions from Plutchik’s model, and an additional emotion (empathetic), which has been shown to be expressed in discussions about chronic illnesses in online health communities. We believe the chosen emotions are related to individuals’ mental well-being and can provide some insights into the public mood during the COVID-19 pandemic. Figure 3 shows the distribution of the chosen emotions from the annotated dataset.

Since a tweet can be annotated with more than one emotion, we create 4 binary classification tasks: annoyed/non-annoyed, anxious/non-anxious, empathetic/non-empathetic and sad/non-sad, to determine if an emotion is expressed in a tweet. To create the negative samples for an emotion category, we sample an equal amount from the other emotion categories. Following this, we shuffle the positive and negative samples and perform an 80/10/10 split, to create the training, validation, and test sets, respectively. Although the number of positive and negative samples is matched

---

1 instructions on accessing data can be found at https://github.com/gitdevqiang/SenWave
and balanced, they are not necessarily the same number of samples in the training, validation, and test splits, thereby representing a more natural distribution of labels. Furthermore, we repeat the negative sampling (with replacement) 10 times for each emotion category, to create 10 dataset samples per emotion category. This aims to reduce any sampling bias in selecting the negative samples. Finally, we have 10 data samples for each emotion with train, validation, and test sets. We train our models with all data samples and report the mean performance across the 10 data samples per emotion category. For all models, we train using binary cross-entropy, loss to align with the target class distribution. We use the Adam optimiser with a learning rate of 5e-5 and set the number of epochs to 4 and batch size to 16 for the final model based on prior empirical experiments.

We compare our model with two general pre-trained models: BERT\(^9\) and XLNet\(^27\). These models, which have achieved state-of-the-art performance on various natural processing tasks, serve as our baselines.

**Results**

Table 1 presents the results obtained using our model, EmoBERT, compared with generally pre-trained language models, i.e. BERT and XLNet. EmoBERT achieves higher performance across all the considered emotion categories. On average, EmoBERT outperforms other state-of-the-art approaches by at least 3% F1 score. This shows that incorporating emotion-specific knowledge in pre-trained language models is effective for detecting fine-grained emotions.

Table 1: Emotion detection results averaged across 10 dataset samples. The numbers are percentages. Best results are in **bold**. Precision - P, Recall - Re and F1 score - F1.

| Model | Annoyed | | | | | | Anxious | | | | | | Empathetic | | | | | | Sad | | | |
|-------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|
|       | Pr     | Re     | F1     | Pr     | Re     | F1     | Pr     | Re     | F1     | Pr     | Re     | F1     | Pr     | Re     | F1     | Pr     | Re     | F1     |
| BERT  | 0.78   | 0.80   | 0.78   | 0.68   | 0.73   | 0.69   | 0.83   | 0.74   | 0.77   | 0.71   | 0.77   | 0.73   |
| XLNet | 0.74   | **0.83** | 0.77   | 0.63   | **0.81** | 0.69   | 0.73   | 0.67   | 0.69   | 0.69   | 0.77   | 0.72   |
| EmoBERT | **0.81** | 0.82   | **0.81** | 0.72   | 0.73   | **0.72** | **0.84** | 0.79   | **0.81** | 0.71   | 0.84   | **0.76** |

**Significance Test**

To determine if our results are statistically significant, we perform a paired T-test to test the differences between the model results for all 10 runs per emotion category. The result confirms that the improved F1 score results from EmoBERT over BERT are statistically significant at \( p < 0.05 \) across all emotions, except for the anxious emotion.

**Tracking Emotional Toll of COVID-19 on Twitter**

To further illustrate the power of our model, we present a focused analysis of the impact of the COVID-19 pandemic on well-being and emotions in London, United Kingdom (UK), using EmoBERT. We study a collection of tweets geo-located in London in March 2020 and compare them to the same period in 2019. The tweets were collected using the new Twitter API for academic research that grants access to the full Twitter archive\(^4\). Initially, we collected all tweets geo-located in the UK and then extracted from that pool tweets that contained London in the full name of the place information\(^5\). Non-English tweets, replies, and re-tweets were not included in the collected tweets. Our resulting sample consists of 361,384 tweets for March 2020 and 352,678 tweets for March 2019. We have selected that period because the lockdown due to COVID-19 was announced by the UK government in March 2020.

We classify each tweet from the respective periods with one of the 4 emotions (annoyed, anxious, empathetic and sad). Since we evaluate our model on 10 different dataset samples, the 10 classifier models for each respective emotion are ensembled and the average of their probabilities is used to predict the final class labels. For all emotions, each tweet is classified as expressing that emotion or not (e.g., sad/non-sad). The emotion expressed in each tweet is measured by the probability of the positive class membership of that emotion. We define the probability of emotion \( p_d \) present in tweets on day \( d \) and use the logistic regression classifier as follows:

\[ p_d = \frac{1}{1 + e^{-\beta}} \]

\[ \beta = \sum_{i=1}^{n} w_i x_i \]

where \( x_i \) are the features of the tweet and \( w_i \) are the weights.

\(^4\)https://developer.twitter.com/en/solutions/academic-research/products-for-researchers

\[ p_d = \frac{1}{1 + e^{\exp(-x_d)}} \]  

(1)

where:

\[ x_d = \frac{1}{N_d} \sum_{t=1}^{N_d} \ln\left( \frac{p_{td}}{1 - p_{td}} \right) \]  

(2)

Where \( x_d \) denotes the proportion of emotion present in tweets on day \( d \), \( N_d \) is the number of tweets on day \( d \) and \( p_{td} \) represents the probability of emotion expressed in tweet \( t \) on day \( d \). We take the mean of the logit function in equation 2, because the means of skewed variables are not necessarily representative of those variables. Probabilities tend to exhibit such skewness, because they are bounded, so it is often cleaner to do algebraic manipulations on an unbounded scale, such as logit, then back-transform\(^4\).

Figure 4 shows the comparisons of the emotions for each respective year. Generally, the proportion of tweets expressing emotions in 2019 show a similar trend over time. There is no notable change in the emotions for that period. Similarly, the tweets expressing annoyed and sad emotions in 2020 are consistent during the period. Interestingly, the tweets expressing the anxious emotion rise sharply after the 5th of March (after the first COVID-19 related death was announced in London)** before levelling off and declining around the 20th of March 2020. The tweets expressing empathetic emotion rise slightly in the middle of the month and begin to decrease after about 10 days.

![Figure 4: The trend of emotions from the 1st March through 31st March for years 2019 and 2020](https://www.bbc.co.uk/news/uk-england-london-56271001)

Although the trend in tweets expressing annoyed and sad emotions over time is similar in 2019 and 2020, the level of emotions is slightly higher in 2020 than in 2019.

We measure the effect size using Cohen’s \( d \) to determine the difference between each respective year’s emotions\(^2\) and report \( p\)-value after Benjamini-Hochberg p-correction. On average, more tweets expressed annoyed (Cohens \( d = 0.21 \), \( p = 0.05 \)), anxious (\( d = 1.99 \), \( p < 0.05 \)), empathetic (\( d = 0.73 \), \( p < 0.05 \)) and sad (\( d = 0.5 \), \( p < 0.05 \)) emotions in

\( **\) https://www.bbc.co.uk/news/uk-england-london-56271001
2020, compared to 2019, suggesting that there is a difference in the expression of emotions in tweets as a result of the COVID-19 pandemic.

To understand more deeply the distribution of emotions in tweets related to the COVID-19 pandemic, we further extract, from the tweets geo-located in London, those containing the following keywords: coronavirus, corona, covid, covid-19, coronaoutbreak, 2019nCoV, pandemic, epidemic, wuhan and lockdown and analyse them in-depth. We build a matrix of the emotion distributions with the most frequent hashtags that appear in the filtered tweets. Hashtags are strong indicators used to provide context, emotions or topics related to a tweet. Figure 5 shows the heatmap of the distribution of emotions and the top 40 hashtags. The distribution of emotions is calculated as the number of instances in which the emotion is expressed in a tweet with the hashtag, divided by the total number of tweets containing that hashtag from the body of extracted tweets\(^2\). As can be seen in Figure 5, tweets expressing sad and anxious emotions are prevalent in the hashtags used in the COVID-19-related tweets. Hashtags in tweets expressing empathetic emotions are common in campaigns about staying safe and staying at home, while hashtags in tweets expressing annoyed emotions are more related to the crisis and chaos in the UK, as a result of the pandemic.

The majority of the hashtags used in tweets expressing the sad emotion are #isolation, #quarantine, #londonlockdown and #coronapocalypse. London is a travel hub and, understandably, the tweets are expressing sadness about travel restrictions, such as quarantine, isolation and lockdown in London. The tweets also express sadness about what some people have regarded as the 'end of the world' because of the pandemic. Amongst tweets with the #coronavirusoutbreak, #covid-19, #coronavirusupdates and #coronaoutbreak hashtags, the anxious emotion is preponderant. This suggests that a significant amount of tweets are expressing anxiety about the COVID-19 pandemic and particularly around the updates provided on the virus.

![Figure 5: Top 40 hashtags for COVID-19 related London tweets in March 2020](image)

The majority of tweets with hashtags such as #coronavirusuk, #covid-19uk and #coronacrisisuk express the annoyed emotion predominantly. Amongst these, hashtags related to the coronavirus crisis in the UK are most common. It can be seen that #staysafe, #staysafelives and #nhs hashtags, which are related to the UK government official campaign advising people to “stay home, save lives and protect the NHS (National Health Service)”, appear more frequently in tweets expressing the empathetic emotion.

**Conclusion**

In this paper, we propose a new method to analyse the emotional health and well-being of global as well as local populations (here, London), by using a new model called EmoBERT. This model adds emotion representation to the previously cutting-edge BERT model, outperforming the current state-of-the-art. Additionally, we have selected emotions related to health and health communities, to showcase a methodology for an in-depth comparison between social media emotions before and during the COVID-19 pandemic. We further showed how these selected emotions can be used to understand the individual topics that are likely to evoke them. We applied the latter locally, on a major location, London, as we expected it to be strongly affected - as confirmed by our analysis. Similarly, for further work, other specific areas can be analysed with the methodology we have defined. Limitations of our approach include analysing only one (freely available) social media platform (Twitter), which may only reflect the emotions of a specific section of the population. Such results could further be extrapolated with other social media sites e.g. Facebook or Weibo, as well as data from the World Health Organisaton (WHO) or local health authorities, for a higher accuracy.
References

6. Avery EJ. Public information officers’ social media monitoring during the Zika virus crisis, a global health threat surrounded by public uncertainty. Public Relations Review. 2017 Sep 1;43(3):468-76.
18. Hayati SA, Muis AO. Analyzing incorporation of emotion in emoji prediction. InProceedings of the Tenth Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis 2019 Jun (pp. 91-99).
Expressing and Executing Informed Consent Permissions Using SWRL: The All of Us Use Case

Muhammad Amith, MS, PhD\textsuperscript{1\*}, Marcelline R. Harris, PhD, RN, FACMI\textsuperscript{2\+}, Cooper Stansbury, MS\textsuperscript{2}, Kathleen Ford, MM\textsuperscript{2}, Frank J. Manion, PhD, FAMIA\textsuperscript{3}, Cui Tao, PhD, FACMI\textsuperscript{1}\textsuperscript{*}

\textsuperscript{1}School of Biomedical Informatics, University of Texas Health Science Center, Houston, TX; \textsuperscript{2}University of Michigan, Ann Arbor MI; \textsuperscript{3}Melax Technologies, Houston TX

\*contributed equally to this work
\*corresponding author:cui.tao@uth.tmc.edu

Abstract

The informed consent process is a complicated procedure involving permissions as well a variety of entities and actions. In this paper, we discuss the use of Semantic Web Rule Language (SWRL) to further extend the Informed Consent Ontology (ICO) to allow for semantic machine-based reasoning to manage and generate important permission-based information that can later be viewed by stakeholders. We present four use cases of permissions from the All of Us informed consent document and translate these permissions into SWRL expressions to extend and operationalize ICO. Our efforts show how SWRL is able to infer some of the implicit information based on the defined rules, and demonstrate the utility of ICO through the use of SWRL extensions. Future work will include developing formal and generalized rules and expressing permissions from the entire document, as well as working towards integrating ICO into software systems to enhance the semantic representation of informed consent for biomedical research.

Introduction

Informed consent (IC) is a process intended to ensure that an individual (or their legally authorized representative), has sufficient information and understanding to voluntarily make a decision about participating in a research study. IC documents provide individuals with the information needed to make a decision about whether to volunteer for a research study and serve as a record of the decisions made during IC. In this way, IC documents serve as an important communication vehicle between the research team and potential study participants (or legally authorized representatives). IC documents provide information about actions the signer of the IC document prescribes as allowable, given the information within the IC document.

The specific research context we address are studies that aim to collect and share biospecimens, the data derived from analysis of those specimens, and other sources of data that can be associated with the study participant. Informed consent documents, when signed, establish and preserve important linkages among the persons signing the consent form, the research team, stewards and managers of specimen and data repositories, and other potential future users of specimens and/or data; the IC documents become a ‘source of truth’ regarding the allowability of potential actions by the research team and others.

There are several notable efforts that focus on open, computable representations of permissions expressed in IC consent documents. For example, the Community Based Collaborative Care Resource Work Group within Health Level Seven (HL7) is developing a consent resource specification \url{https://www.hl7.org/fhir/consent.html}. The resource targets four use cases: (1) privacy consent directive, (2) medical treatment consent directive, (3) research consent directive, and (4) advance care directives. Currently, the published version (v4.01) is limited to the privacy consent directive. Of particular relevance to our interest, developers of the FHIR resource define a consent directive as “the legal record of a patient's (e.g. a healthcare consumer) agreement with a party responsible for enforcing the patient's choices, which permits or denies identified actors or roles to perform actions affecting the patient within a given context for specific purposes and periods of time.” A consent form is defined as “Human readable consent content describing one or more actions impacting the grantor for which the grantee would be authorized or prohibited from performing. It includes the terms, rules, and conditions pertaining to the authorization or restrictions, such as effective time, applicability or scope, purposes of use, obligations and prohibitions to which the grantee must comply.” The FHIR consent resource does not yet address the challenge of expressing the multiple consent directives that may be present in a single IC document, nor the challenge of linking detailed information such as actors, actions, and...
conditions to one or more consent directives within the document. Our teams have reviewed thousands of research and clinical IC documents, using both manual and machine learning based approaches. Many if not most of the documents we reviewed include multiple permission-directives (i.e., consent directives), are often stated ambiguously, and contain relations to clauses throughout the document that are necessary to understand the permission itself. Since the FHIR specification is not fully developed, value sets that reflect the full spectrum of terminology needs related to research permissions are not yet identified.

Another example of efforts to support computable representations of permissions in IC documents is published by the Global Alliance for Genomics and Health (GA4GH). Included in their regulatory and ethics toolkit are suggested consent clauses for genomic research and for rare disease research, as well as machine readable consent guidance. The consent clauses are intended to be used as a resource for researchers as they draft IC documents, with language that is consistent with the GA4GH standards for essential consent elements. Currently, the consent clauses are available only as text-based sentences; there is no machine-processable format. The GA4GH machine readable consent guidance similarly provides another resource for researchers as they construct IC documents, however this resource provides a mapping from a set of sample consent clauses to terms that have been adopted by the Data Use Ontology (DUO). DUO is published within the OBO Foundry suite of ontologies, and based on the Basic Formal Ontology that is the upper level ontology on which all OBO Foundry ontologies are built [http://www.obofoundry.org/ontology/duo.html](http://www.obofoundry.org/ontology/duo.html). The intent of DUO is to standardize downstream data use concepts into a machine-readable format while tagging datasets with restrictions about their usage; DUO lacks detail about the IC process and documents.

While the HL7 and GA4GH efforts are intended to support consistent and comparable exchange of information about consent directives and potentially more detailed information about permitted actors, actions, and conditions, neither supports full semantic interoperability including IC processes and entities expressed in IC documents. A desirable feature of full semantic interoperability is that it allows for the use of semantic reasoners, based on predicate logic, to infer relationships between information; in this case, information about permissions derived from the informed consent documents.

In this study, we build on our previously developed Informed Consent Ontology (ICO) [https://github.com/ICO-ontology/ICO](https://github.com/ICO-ontology/ICO). Intended for use as a reference ontology, ICO represents processes and information entities pertaining to informed consent in biomedical investigations. ICO is based on the upper level Basic Formal Ontology (BFO), and adheres to the OBO Foundry framework for authoring and editing BFO-based ontologies; it is licensed under the [Creative Commons Attribution 4.0 International Public License](https://creativecommons.org/licenses/by/4.0). The August 2020 ICO release includes a full alignment with DUO and a new set of ‘information content entities’ that reflect categories of information needed to fully express information about permissions, i.e. permission-directives as well as relevant-to-permission information. The two classes of ‘directive information content entity’ and ‘designative information content entity’ are hierarchical, with subclasses of ‘permission directive’ and ‘permission condition directive’, and ‘designated actor’, ‘designated action’ and ‘designated object’. Each of the subclasses has a small number of additional hierarchically modeled classes. We define a ‘permission-directive’ as a directive information content entity that prescribes an allowable action, where that action is otherwise impermissible. Permission directives are usually expressed in IC documents as statement(s) that, upon signature of the consent form, authorizes actions by actors that would otherwise not be allowable. Phrases such as “I consent to”, or “I agree to” may be indications of permission-directives. More specific information about what action may occur, when it may occur, why it will occur, and any conditions under which it may occur is commonly found in other clauses, and often not co-located with the permission-directive.

**Semantic Web Rule Language.** Semantic Web Rules Language (SWRL) is a W3C semantic web language extension to implement predicate logic on variables. SWRL’s benefit in relation to ontologies is to enable expressive machine-based rule reasoning for instance-level data and add additional expressiveness on top of the Web Ontology Language (OWL). The composition of SWRL contains two conditional propositions - antecedent condition (AC) and consequent condition (CC). The AC is a conditional body proposition that describes the "if" aspect of a rule statement and the CC is the head proposition that describes the "then" condition. Basically, the completeness of a SWRL is contingent on whether the conditions described in the body (AC) and the head (CC) conditions are satisfied. The expression below shows an example of a simple SWRL statement.

```
Person(?a) ^ of_age (?a, ?age) ^ swrlb:greaterThan(?age,18) ^ residesInCountry(?a, United States) -> Adult(?a)
```
For the simplified example described, a rule statement for determining legal adulthood is complete if only if the entity (?a) is a person that has an age greater than or equal to the value of 18 and resides in the United States. Each component within the statement is an atom which is the core building block of the statement. These SWRL atoms describe the classes (Person), individuals (a, age), data values (18), and other built in utility arguments. For this paper we introduce the possibility of expressing the complexity and execution of permissions in IC documents using SWRL and the ICO.

Research Objective: In an effort to extend the ICO, we intend to use SWRL to enhance the expression and application of informed consent. The application of SWRL could further enhance the expressive power of ICO by mapping specific data entities in the informed consent process (actors, data, agreement, etc.) and operationalize it for the software system. We presume that ICO through the extension of SWRL can express and operationalize permissions embedded in informed consent documents. The outcome of this study would demonstrate the practicality of using SWRL to extend the operational use of ICO for software systems.

Methods

To determine if our approach could successfully represent contemporary permission constructs we examined a single IC document, the sample primary IC document from the NIH “All of Us” research program, dated June 20, 2018. The All of Us program seeks to recruit one million or more diverse participants to a large observational trial with the end goal of supporting a variety of use cases within precision medicine (https://allofus.nih.gov/). When signed, an “All of Us” IC document authorizes the collection, retrieval, sharing, and reuse of participant’s basic data, health data, physical measurements, biological specimens, fitness trackers and additional data from sources such as health registries, pharmacies, and claims data. Content from electronic health records can be used if the HIPAA authorization form is signed. We selected this IC document for this study because of the scope and familiarity of the “All of Us” study, and because in our earlier review of IC documents retrieved from ClinicalTrials.gov, the readability of the “All of Us” document was at an easy to read level, and thus did not present our study with ambiguous language. We did not include the separate HIPAA authorization form that is required if participants choose to allow access to their electronic health record data.

The All of Us IC document includes 270 sentences. Only one was a permission-directive, 130 (48%) were relevant-to-permission sentences, and 139 (52%) sentences categorized as “other,” i.e. important information but not relevant to the purpose of this study. A single permission-directive was identified: “I freely and willingly choose to take part in the All of Us Research Program.” Examples of permission-relevant sentences included the following:

- “These are results that could be used by a healthcare provider to take better care of you. For example, if any of your physical measurements are outside of what we would expect, we will tell you so you can follow-up with your healthcare provider.”
- “If I give a blood, urine, or saliva sample, it will be stored at the All of Us biobank. This includes my DNA.”
- “Information that researchers learn by studying my samples will be stored in the All of Us databases.”
- "Researchers will do studies using the All of Us databases and biobank. Their research may be on nearly any topic."

With each permission there are one or more consent directives that link the designated instance data for designated actors, designated actions, designated purposes, and designated objects. Essentially a data instance of a consent directive would link to instance data, and related entities from the consent document. As we will show later, through semantic rules, additional links based on the few links described will be inferred to show implications of agreeing to the consent. Definition 1 of our model elucidates the notion that IC documents may contain various permissions and that each permission has a consent directive entity. Definition 2 further elaborates on our model with designated entities of actors, action, purpose, and object.

Definition 1. For any informed consent form icf there exist a set of permissions P belonging to an informed consent form icf. Also, there exists one consent directive cd for every permission P.

\[
\{P_1, P_2, P_3, \ldots, P_n\} \in icf
\]

\[
cd_n \in P_n \text{ where } n > 0
\]

Definition 2. We define designated entities DE as a set of designated action da, designated actor dr, designated purpose dp, and designated object do, that has a relation ret about with a consent directive cd that describe A.

\[
DE = \{da, dr, dp, do\}
\]

\[
ret_{about}(cd_n, DE_m) = A; \text{ where } n > 0 \text{ and } m > 0
\]
Each permission describes probable designated actors that include the subject (the person signing the consent form – the designated permitting actor), the institutions and organization that may be involved – e.g., National Institute of Health, some research organization, All of Us initiative, and the research team members including the principal investigator (the designated permitted actors). Additionally, designated object for each of the models are described if the permission refers to subject’s data – e.g., general health information, personal health information, or some biospecimen. The models also include the designated purpose and designated action that expresses motivation (research, therapy, etc.) and the general action on the designated object for which permission is being given (disclosure of data, data collection, etc.).

The permission includes planned processes that are described in the informed consent ontology where the designated entities link either explicitly or implicitly (through machine inference). For the former and depending on the expression, there is a defined implication that consenting is act of authorizing some type of permission to another actor. For the implicit links to the planned process, there are allusions to future processes if the subject consents – e.g., granting permission to the researchers to use the shared data, agreeing to inform the patient of change of care, storing the subject’s data, etc. These implicit inferences, since the entities are modeled as instance data, will be derived from Semantic Web Rule Languages (SWRL) to be discussed in the next section. Definition 3 explains that for a permission, there are predicates (defined as A) that are members of a permission of an informed consent. These predicates make references to planned process(es), which can be inferred.

**Definition 3.** Every permission P of an informed consent form entails a subset A that produces an allusion → to some planned process pp. Essentially, any derivative of A is an allusion → to some planned process pp.

\[
P \ni \{\{A_1, A_2, \ldots, A_n \rightarrow pp\}\}
\]

where \(A \iff rel^{about}(cd, \{da_m, dr_m, dp_m, do_m\}), m > 0\)

\(\vdash rel^{about}(cd, \{da_m, dr_m, dp_m, do_m\}) \rightarrow pp\)

With our general defined model of abstracting permissions from informed consent documents, we developed SWRL expressions that can be reasoned on the available entity data extracted from the permissions to infer entities that may be implied when agreeing to participate in a study. We chose the aforementioned four permissions from All of US document to test our approach.

**Results**

For each of the aforementioned examples (Use cases 1 through 4), we produced individual SWRL expressions to generate implied agreement for procedures based on entity data derived from the informed consent document. We populated the ICO ontology with instance level data associated with probable entities in Protégé 5.5. We executed the SWRL rules and reviewed the generated inferences to assess execution of the rule in an iterative process to refinement. We utilized the built-in Hermit reasoner of Protégé which supports the generation of inferred axioms from SWRL.

**Use Case 1.** This use case involves understanding the process of participating in the informed consent process for collection of participant’s health data (designated process of collection and designated health information) for research (designated biomedical research purpose). The underlying assumption that once the subject agrees to this permission (consent directive) there is presumed permission (legally effective consenting) and that the researchers have informed the participant of the study process (explaining to participant candidate in informed consent process) and met the baseline requirements for informing, which are inferred and reasoned through the ontology. Following the participant (designated permitting actor) providing consent, the research actors become designated permitted actors by inference. Figure 1 models Use Case 1 permission, and Code Listing 1 shows the SWRL expression to execute our use case.
Please check the box below if you agree to take part: I have read this consent form (or someone read it to me). I understand the information in this form. All of my questions have been answered. I freely and willingly choose to take part in the All of Us Research Program.

**Figure 1.** Use Case 1’s instance-level ontological model

Listing 1. Use Case 1’s SWRL expression.

The SWRL code from Listing 1 successfully executed the embedded rules (See Figure 2). From the figure, it shows:

1. The participant (designated permitting actor), through inference and formally agreeing to the permission, participating in (participates in) the legally effective consenting (legalconsent).
2. The research team consisting of the primary investigator (“pi”) and the team (“team”) are inferred as designated permitted actors as a result of the subject executing consent to the permission. In addition there’s an expectation of participation in (participates in) explaining details of the study as expressed in the informed consent form (see 4).
3. participates in is an object property of ICO that has an inversed version, has participant. As a result, the ontology reasoned has participant for the subject that consented to the permission (“I”) with the legally effective consenting (a planned process) as the domain.
4. Similar to 3, the planned process of explaining to potential participant through study informing process inferred has participant for the research team and the PI (“pi”). There was also a set of planned processes relating to explaining the study information in the informed consent form that was reasoned. Through implication, these covered the informed consent requirements when discussing the study.
Use case 2. The second use case outlines the permitting actor’s agreement (after consent) that, if participant data were to have indications that were abnormal, then the researchers’ responsibility is to inform the participant about results in order that he/she may seek a health care provider; the expectation is that participant will follow through. This use case model infers (participates in) that the consenting permitting actor, is likely to seek standard medical treatment and infers that the researchers, both the primary investigator and the team will carry out (participates in) the informing process (act of informing). Similar to first case, the researchers, upon the subject’s agreement, will be designated permitted actors. Figure 2 visualizes this use case model’s execution of the permission’s agreement by the subject and Listing 2 shows the corresponding SWRL code.

Results that might change your medical care. These are results that could be used by a healthcare provider to take better care of you. For example, if any of your physical measurements are outside of what we would expect, we will tell you so you can follow-up with your healthcare provider. You will have to pay for the cost of follow-up care with your own healthcare provider.

Figure 3. Use Case 2’s instance-level ontological model.

Listing 2. Use Case 2’s SWRL Expression.

From the execution of the SWRL coding, the following figure (Figure 4) shows the various inferences:

1. The designated permitting actor (the subject who agreed to the permission) is inferred to comply in participating in (participates in) following up with medical care if the subject’s data is abnormal.
2. The investigation team of the primary investigator (“pi”) and cohort (“team”) is inferred in participating in informing (“toinform”). This also includes assigning them as designated permitted actor(s).
3. The SWRL execution also includes inferring the inverse of participates in (has participant) for the designated permitted actors and designated permitting actor.

Use Case 3. The third use case describes a model where upon the subject’s execution of agreement, it permits the execution of authorizing (act of authorizing) the sharing of his/her biospecimen data (act of data sharing) and the storage of subject’s biospecimen data (act of storing a specimen) by the inferred designated permitted actors of the primary investigator and his/her team. The data to be stored is restricted (via restricted to) just to the urine, blood, and saliva. This use case’s SWRL code is described in Listing 3.

Listing 3. Use Case 3’s SWRL expression

The results of SWRL expression from Listing 3 are presented in the figure below (Figure 6). As a result of the research subject executing consent here are the inferred expectations:
1. The research subject who consented to sharing data participates in act of data sharing (“tosharedata”).
2. The subject authorization is restricted to storing and sharing of his data – “tostoredata” and “tosharedata”.
3. Research team members involved are inferred to be designated permitted actor(s) and are inferred with the expectation to store the subject’s data with participates in.

4. From the participates in inverse feature, the ontology infers participant for act of data sharing (for the subject “I”)...

5. …and for act of storing a specimen (for the “pi” and “team” — researchers). Additionally, the storage of data is restricted to just the participant’s saliva, blood, and urine with an inference of is restricted to.

Figure 6. Series of screenshot showing inferences generated through SWRL for Use Case 3.

Use case 4. The model for this use case expresses conditions that indicate restrictions based on the data available from the participant (urine, blood, saliva of type designated biospecimen) and their use will be for any research topic in the future (designated process of unspecified use and designated research purpose). Structurally, the abstraction for this model shares some similarity with Use Case 3. If the research participant consents, then the permitting actor authorizes (act of authorizing) the use of the data by the researchers (inferred as designated permitted actors) is restricted to (by inference) the subject’s urine, blood, and saliva. The code listing (Code Listing 4) shares some similarity with the previous that was executed.

Figure 7. Use Case 4’s instance level ontological model.
Listing 4. Use Case 4’s SWRL expression

The following figure (Figure 8) shows the successful execution of the use case’s model from Listing 4. In the figure we show:

1. The act of authorizing by the subject is restricted to the act of using study participant data.
2. Similar to the other example use cases, the research team is inferred to be designated permitted actor(s) and they will participate in using the data (act of data using study participant data).
3. Like Use Case 3, the act of using study participant data is restricted to subject’s blood, urine, and saliva, and through the inverse of participates in, expresses the inference that this act involves the research team of the primary investigator (“pi”) and his/her research team (“team”)

Figure 8. Series of screenshots showing inferences generated through SWRL for Use Case 4.

Discussion

In this project, we demonstrate the utilization of the Semantic Web Rule Language (SWRL), an extension of OWL, to provide inferencing of entity instance data from informed consent documents. The use of SWRL could effectively enhance the application of ICO to manage complex and essential IC data that must be linked for understanding permissions. Moreover, this work could be potentially integrated into software systems. In this way, we address the gap in establishing traceability from details included in the consent form to repositories and other systems.

Figure 9. Software system for managing and querying informed consent showing the integration.

While much of what we presented was implemented through the Protégé environment, with the SWRLAPI combined with the ICO model, one can potentially create custom software integration to be incorporated into software systems, as shown in Figure 9. The study participant, through an interface, confirms his/her agreement to participate in a study by signing an IC document. The data from the IC document, along with the contextual and meta-data information that are not explicitly stated, are transferred from the software interface to the system. The system software converts these data to instance data. This instance data is added and labeled to ICO, and with the assistance of SWRL extensions, reasoning is performed on the data. The consent data (including the reasoned consent data) can later be queried by clinicians.

From our four use cases we demonstrated the utility of our proposed method. Each use case demonstrated not only translating the permission to a computable format using ontology language (OWL and SWRL), but they also demonstrated predictability inferencing, using ontology-based reasoning to generate links to entities that may not be explicitly evoked in the consent form (e.g., identifying designated actors).

Several limitations are noted. The rules and the ontological translations are only examples, and used to show proof of concept; they are not comprehensive or complete since we limited our sample to one IC document with one permission-directive. In addition, we have only used the “is about” relation; our goal here was to identify information content that could be tagged and serve as a foundation for use in computable semantics such as inferencing operations.
There were a handful of concepts that could not be precisely expressed, due to a lack of coverage of minor entities. Some of this apparent lack of coverage in ICO may result from the need to import content from other BFO-based ontologies in the OBO Foundry. These cases need to be enumerated and then used to either identify gaps in ICO itself, or negotiate inclusion in other ontologies in the OBO Foundry. ICO provides a general model of informed consent and does not include in its scope the modeling of regulatory concepts; there is a need to support the creation of SWRL rules to help extend legal, regulatory, and policy expressions. One probable future goal is to automate the translation of SWRL rules from informed consent documents. However, a more achievable and transparent method would be a visual authoring tool to create the SWRL expressions that align with the permissions in the informed consent document. We are encouraged from our experience in authoring the expressions based on the ontology and the SWRL rules, that this could be generalized for most of the informed consent documents as there are consistent semantic structures involved in accurately expressing permissions. Our group is pursuing the creation and validation of such models. Finally, the focus of the work we presented is permissions embedded in the informed consent document. Our observation in the use cases are of similar and repeated abstractions, and contexts that were on the document or meta-level but not explicitly stated on the permission level. Part of our future direction is to approach the ontological expression on a "document level" that would collapse or merge the abstraction to avoid repetition and redundancy.

Conclusion

Through the use of semantic web rule language (SWRL) that enables machine-level reasoning on the instance-level data of ICO, we were able to demonstrate our proof of concept method to link and generate information from entities derived from four permissions in the All of Us informed consent document. Our efforts allowed us to extend the application use of ICO to be potentially integrated with software systems to manage complex authorization information that can later be supported by queries. Our future direction is to work towards formalizing a set of generalized rule extensions for ICO to accurately express document-level permission entities from informed consent.

Acknowledgements. This research was supported by the NIH/NCI under Contract Number 75N91020C00017 (Manion PI), and builds off work supported by an award from the Michigan Institute for Data Science (Harris PI), and NIH/NHGRI U01 HG009454 (Tao PI); and by the National Institute of Allergy and Infectious Diseases of the National Institutes of Health under Award No, R01AI130460 (Tao, PI) and R01AI130460-03S1 (Tao, PI). The University of Texas Health Science Center at Houston has research-related financial interests in Melax Technologies, Inc.

References

Children’s Designs for the Future of Telehealth

Erin Beneteau, MS¹, Ann Paradiso², Wanda Pratt, PhD¹
¹University of Washington, Seattle, WA; ²Microsoft Research, Redmond, WA

Abstract

Telehealth has increased dramatically with COVID-19. However, current telehealth systems are designed for able-bodied adults, rather than for pediatric populations or for people with disabilities. Using a design scenario of a child with a communication disability who needs to access telehealth services, we explore children’s ideas of the future of telehealth technology. We analyzed designs generated by six children and found three provocative overarching design themes. The designs highlight how improving accessibility, accommodating communication preferences, and incorporating home based sensor technologies have the potential to improve telehealth for both pediatric patients and their physicians. We discuss how these themes can be incorporated into practical telehealth designs to serve a variety of patient populations—including adults, children, and people with disabilities.

Introduction

The design of telehealth technologies for the pediatric population is currently an under-explored area, as is the design of telehealth systems for children with disabilities. Children’s ideas for technology design can provide unique insights, particularly when designing technologies geared towards children as primary users. Concepts such as inclusive design have also shown how designing for disabled users can benefit people with a range of abilities. Our study explores how telehealth can better serve the pediatric population including, but not limited to, pediatric patients with disabilities. This paper describes how designs created by children, using the design prompt of a disabled pediatric patient, can inform the development of future telehealth technologies.

Telehealth and telemedicine incorporate a wide variety of platforms and services spanning synchronous, online video visits to asynchronous electronic health records (EHR). The majority of telehealth systems assume that users are able-bodied adults. However, designing health technologies for the able-bodied alone can perpetuate healthcare inequities. Moreover, with the increased use of telehealth as a result of the COVID-19 pandemic, we question how telehealth systems geared towards adult users can be designed to better support children’s engagement with their primary healthcare provider. In a recent opinion piece in the Journal of the American Medical Association, pediatricians called for the medical community to seize this defining moment in history and to reconsider how telehealth can best serve the pediatric population. Drawing on user-centered design principles, we chose to ask the potential users themselves, children, for their ideas of the future of telehealth.

We conducted one-on-one, online co-design sessions with six children between the ages of 9 and 13. Our findings revealed three provocative design themes: (1) drones, robots and sensors to the rescue: expanding the scope of telehealth, (2) health data is controlled by the family: sharing with clinicians is optional, (3) beyond talking heads: supporting diverse communication needs. In this paper, we explore what these design themes entail and how they can be practically applied to existing and near-future telehealth technologies. We also discuss how building on these themes can benefit a wide range of telehealth users.

Related Work

A historic review of the literature shows that the role of children in medical communication interactions has been underexplored, with the medical research community frequently undervaluing the contributions children can make to their medical care. Thankfully, this historic perspective in pediatric medical research has begun to change, particularly in the field of medical informatics. The perspectives of children and adolescents on their healthcare communication, information sharing and decision making has been recently explored through qualitative interviews, participatory design, and user studies.

Garth et al. explicitly investigated the perceptions of healthcare communication and partnership of pediatricians, parents, and children with motor, mobility and communication disabilities. The authors found that pediatricians and parents viewed children as important contributors to the healthcare partnership. Pediatricians made explicit efforts to engage with pediatric patients to encourage the child’s participation, including talking about topics of interest to the child and being at the same eye level as the child. While this study is specific to in-person healthcare
visits rather than telehealth, the concepts of including the pediatric patient as a partner in healthcare and ensuring that communication is tailored to the child’s level, are important themes that can be applied to pediatric telehealth.

Co-design with children for technology and disability

Participatory design and co-design methods are commonly used in the Human Computer Interaction (HCI) and Computer Supported Cooperative Work (CSCW) research communities to generate insights on people’s needs and ideas to support those needs. These methods are particularly useful when designing technologies that may be used by children. Druin states that “the better we can understand children as people and users of new technologies, the better we can serve their needs”.

A frequent approach to co-design with children is to begin with a story which establishes the design problem. Light, et al. created a story in which they described a child who was unable to use their voice to communicate and who required some form of augmentative and alternative communication (AAC) technology. Light’s team recruited six children to generate designs of AAC systems. The children created designs which were not constrained to communication alone, instead they incorporated multiple functionalities. The design features generated by the children included functionality that is now pervasive in mainstream smartphone technologies such as the ability to play music, access email, play games, integrate popular themes, and have text to speech output. At the time of the study, these features were not ubiquitous as an all-in-one technology. Light et al.’s work demonstrates how presenting children with a design problem scoped to meet the needs of a child with a disability can lead to design concepts which can ultimately be used by a wide range of users.

To engage children who have chronic health conditions or complex disabilities, researchers need to consider the stress and burdens the co-design procedure may place on participants, particularly when participants are minors. In their design of a support system for children with cancer, Ruland et al. chose to recruit healthy children for early stage co-design sessions, involving children with cancer at later stages of the design and development process. Similarly, Light et al.’s team did not include children with disabilities in their early-stage exploratory design work but focused the design prompt on technology for a child with communication disabilities.

Building on the co-design methods used in prior work, we chose to explore the area of pediatric telehealth using a scenario of a child who is not able to use their voice to communicate, and instead uses an AAC device. By choosing this scenario, we investigate telehealth design “from the margins” instead of designing for the “typical” able-bodied, adult. In our study, we learn how designing for this specific scenario leads to provocative, creative designs that have the potential to be practical and useful to a variety of telehealth users.

Methods

Our study design is inspired by Light et al.’s work and based on concepts by Druin, in which child co-design participants are considered experts in their knowledge domain and treated as equals to adults in the design process. Due to COVID-19, we conducted our co-design sessions over video conferencing in one-on-one sessions with six children. Our co-design protocol was carefully crafted to consider the ethical implications of working with children. Our study was reviewed and approved by the sponsoring organization’s institutional review board.

Participants

We used purposeful sampling of existing networks to identify children who could engage in a video conference session for approximately one hour in length. Due to the inherent abilities required when conducting co-design sessions over video conferencing, we did not recruit any children with disabilities for this initial, exploratory study. Parents were emailed both a consent and an assent form. Parents were instructed to review the assent form with their child. Signed consent forms were obtained from the parents of all participant children prior to scheduling the co-design sessions.

We contacted 10 families and had six children volunteer to participate within the study’s timeframe. To maintain diversity in our participant sample, we only included one child participant per family. To protect participants’ identities, we provide generalized demographic information. Participants ranged from age 9 to age 13. Three participants identified as male, two as female, and one preferred not to specify. Three participants identified as white, two identified as white and Asian, one identified as Asian. Participants had family members working in a variety of settings including healthcare, education, transportation, and technology industries.
Procedure

To support children as experts and equal design partners, parents were asked to not participate or be present during the co-design process. The sessions began with the researcher reviewing the assent form with the child, highlighting the purpose of the research and that the child could stop at any time. All six children verbally stated that they wanted to participate. The researcher began by sharing a story of a child named Pat, who was unable to communicate with their voice and used a computer to speak. Pat became sick but could not go to the doctor’s office and the doctor could not go to Pat’s home. Participants were asked to finish the story by describing how Pat was able to get help from the doctor without seeing the doctor in-person.

Both the researcher and the child participant had paper and writing materials available to them in their separate locations and they each independently spent between 5 and 10 minutes of quiet time to sketch and draw their initial ideas of how the story would finish. In this way, the study was designed so that both the child and researcher were considered as equals, engaging in the same tasks. After the quiet time, the researcher asked the child if they wanted to share their story ending first or if they wanted the researcher to share their story first. The majority of children wanted to share their story ending first. If asked by the participant, the researcher shared their own story ending, which was intentionally created to model a design that seemed far-fetched—such as the doctor having an extendable stethoscope.

After the initial story endings were shared, the researcher prompted children for further ideas and additional sketches. Prompts included statements such as “let’s get really out-there and think of some sci-fi options,” “how would the doctor know what Pat’s health was like?” and “what if Pat isn’t able to walk or use their hands? What would we create then?” At times, the researcher would expand on the child’s ideas and ask the child if they liked the expanded idea or would like it to be different. In this way, the co-design sessions became a cooperative exercise between researcher and participant, discussing ideas and building on them, with the researcher always following the child’s lead.

Analysis

Co-design sessions were recorded for review and analysis. During the co-design sessions, the researcher took notes in real-time and wrote memos, synthesizing the co-design sessions. The first author reviewed recordings and photographs of the designs, using an inductive approach to compare, contrast, code and categorize design elements. All authors reviewed images of designs placed in the initial categories and discussed the design concepts, ultimately refining the categories into three broad themes.

Summary of Designs and Thematic Design Elements

Participants generated multiple designs during each co-design session. Initially, participants generated ideas using existing technologies, such as using video conferencing, text messaging, or emailing. Our analysis focuses on designs that children generated after prompts to expand on existing technologies, such as thinking of “sci-fi” options, ideas that “didn’t need to be real,” or additional design constraints such as “what if Pat can’t use their hands to type?” Based on our analysis, we found three conceptual design themes which we describe in detail, below.

Theme One: Drones, robots and sensors to the rescue: Expanding the scope of telehealth

In this theme, participants designed ways for the physician, patient and/or patient’s parent(s) to collect health data that is normally obtained during an office visit. Solutions focused on: (1) transporting and using medical equipment and sensor technologies, (2) providing a better view of the patient than traditional video conferencing solutions, and (3) using health sensor data.

Transporting and using medical equipment and sensor technologies

Our participants created a variety of solutions for collecting health data in the home environment including various forms of “sending” standard office visit medical equipment to the patient’s home. Participants who created solutions where equipment is sent to the home emphasized the need for it to be easily cleaned and used by the child and/or their parent(s). For example, one participant designed a drone (Figure 1) that could have sanitized equipment placed inside it (such as a thermometer, oxygen measure, blood pressure monitor, etc.). The participant explained that the doctor’s office would pack the equipment within the body of the drone and the drone would fly to the patient’s home. As part of the equipment package, disinfectant wipes were included, particularly for items that would be returned to the doctor’s office. The drone would be clearly marked with red crosses to indicate it was a medical drone. The drone would also have a speaker through which the doctor could announce the drone’s arrival. The drone also was equipped with a camera, so that the doctor could pilot the drone around the patient to obtain a 360° view of the patient during the exam as a solution to overcoming one of the limitations of video calling.
Providing a better view of the patient than traditional video conferencing solutions

Children addressed the issue of the doctor needing a comprehensive view of the patient in a variety of ways. In addition to the idea of the drone with the camera, another participant addressed the challenge of the doctor being able to examine the patient remotely by designing a device which the patient could use to create a video of themselves (Figure 2). In this participant’s scenario, the patient is able to use the handheld device to take photos and video of their injury and could toggle into an “x-ray mode” so that the doctor could get the best images possible of the patient’s injuries. In addition, the patient is able to narrate the video and photos they take using their augmentative communication device. In both the instance of the drone camera and the photo/video recording device, we see that a portable, movable camera is at the crux of the design solution for providing the doctor with a more comprehensive “view” of the patient.

As an alternative to “real” images of the patient, one participant created an interactive three-dimensional, rotatable drawing of a body (Figure 3). This tool would allow either the child or doctor to visually highlight areas of the body that they were discussing, overcoming the current “view” limitations of telehealth video conferencing systems.

Another participant expanded the concept of home-health visits to include virtual reality and robotics (Figure 4)*. The participant created two options in which an in-person visit could be conducted in the patient’s home, without the doctor being physically present. In the first option, the physician controls a robot using virtual reality, so the physician has an “in-person” view of the patient. In the second option, the robot itself would be able to perform basic medical exams and basic diagnoses (such as temperature and blood pressure). The participant noted that in both designs, safety and accuracy were primary concerns and that the robot’s actions should be checked by humans if there are any

---

* This participant declined to share their hand-drawn sketches created during the co-design session and preferred to share digital concept sketches of their designs. Figures 4 & 10 are screenshots of their digital concept sketches.
concerns, including the use of a triage system for “option 2” in which the human doctor reviews the robot doctor’s diagnosis and exam procedures. The home visit robot would provide additional safety protections both for the patient and for the doctor during a pandemic. Both robot options designed by the participant are similar in concept to an existing prototype robot designed for basic inpatient services. However, unlike the prototype, the participant’s designs are geared towards telehealth medical care which can be conducted in the patient’s home.

Using health sensor data

During the co-design sessions, participants were prompted with variations of the question: “how would the doctor get the same kind of health information that they normally get during an office visit, such as temperature?” In response, our participants included traditional “office visit” equipment such as thermometers and stethoscopes in their designs in addition to creating apps or other technologies which could measure basic health data through attachments to a smartphone or tablet (Figure 5). One participant created a special sensor that could be placed on the chest of the patient to measure a variety of biosensor data all at once, which could be easily taken off once the health data was collected (Figure 6). Another participant explicitly combined the concepts of a “thought reader” with sensor technologies to create an all-in-one communication and medical exam wearable device (Figure 7). Initially, this participant designed an apron with embedded sensors which could easily be taken on and off. They continued to iterate on the idea of a wearable which resulted in a jacket with embedded sensors to measure respiration, temperature, blood pressure, and other health data. Addressing our design scenario of a child who uses technology to communicate, the jacket includes a hood, which can read the child’s thoughts and transmit those thoughts through a wire to the doctor’s computer. The hood can be easily put on and taken off by the child, so that they have control of when they communicate with the doctor. This design, like so many others from our participants, incorporates user-controlled privacy options through the inherent design of the technology, in addition to combining health sensing data.

Theme Two: Health data is controlled by the family: Sharing with clinicians is optional

As we see with the health sensor designs, participants’ designs implicitly had privacy controls with the ability for the sensors to be easily removed by the user (or the user’s parent). While some participants explained that the child’s parent(s) would use the equipment, other participants indicated that the child might be able to use some of the equipment independently. Ultimately, the designs were flexible enough that the choice of who would use the equipment was left to the family.

Participants discussed different needs for information sharing based on the child’s particular health situation, the child’s abilities, and the child’s health history. A strong, recurring theme between multiple participants’ designs is that the user (child and/or parent) has control of any health data sharing with the physician. Multiple designs across different telehealth platforms feature a “send” button, in which the patient and/or their parent(s) choose to intentionally share health data with the physician (Figures 8-10). One participant developed the concept further to include options for personal health tracking which can be visualized and saved for the patient’s own reference and personal health records (Figure 10). Another participant explained that they included a “send” button in their design (Figure 8) because the child’s parents will know if the child’s temperature is “normal” for their child or not, since some people can have a slightly higher or lower temperature as their “normal” temperature.
Analysis of the participants’ designs reveal an inherent distrust of technology automatically making the judgement on whether to share personal health data with the physician. We see an explicit and intentional design that makes the patient and/or their parent(s) the owners of their data, in control over how they share that data. We also see that designs are created so that the patient and/or their parent(s) are able to obtain the health information data that they want to capture for themselves as a primary feature of the technology. The design assumption is clear: it’s important to capture health data and once the data is captured and reviewed by the child and/or their parent(s), the family members will then decide if it needs to be shared with their medical provider.

**Theme Three:** Beyond talking heads: Supporting diverse communication needs

Our design scenario of a child who cannot use their voice and uses a computer to speak inherently lent itself to design solutions which accommodated many methods of communication. Participants generated design ideas with a focus on direct communication between the pediatric patient and their doctor, in ways that would be easiest for the child to interact. Participants noted that it might take extra time for the child in the scenario to communicate, and as a result, designs incorporate elements that enable faster, easier synchronous communication and asynchronous communication.

One participant modified the design of current video conferencing systems to feature the text-based chat function as the primary visual interface of the telehealth system (Figure 11). The chat shows both the patient’s text as well as the physician’s text in the main window of the screen, similar to many text messaging interfaces. This participant also incorporated a variety of features to enable increased speed and effectiveness of text-based communication through design elements such as: highlighting of important text, word/phrase prediction, speech to text/text to speech, and the use of emojis. The option for sharing video is also available in this design but must be intentionally turned on as a menu item by the patient, instead of being the default focus of the telehealth interaction.

In contrast to existing telehealth systems, our co-design participants focused on telehealth communication methods that would be best utilized by the pediatric patient. One participant explained that they use YouTube for gathering most of their information instead of Google searches, and that they follow different YouTube channels on topics they’re interested in and that they learn of new events and developments by subscribing to those channels. Building on the concept of YouTube, this participant designed a private, asynchronous video chat that the pediatric patient could have with their doctor (Figure 12). The doctor would create short videos to interact with the patient, for example, asking the patient to follow through with a treatment regime or instructions for an upcoming examination. The child would be able to respond to the physician through text-based comments and chats in response to the video. The participant emphasized multiple times that it was extremely important that the videos from the physician concluded with something funny, so that the pediatric patient would watch the full video all the way to the end. We can think of this solution as a pediatric version of a patient portal, in which the pediatric patient is able to asynchronously interact with their care team in a way that is understandable, familiar, and useful to the patient themselves.
Our co-design participants put their design emphasis on ensuring that the pediatric patient’s information needs were met during the telehealth visit. One participant placed additional emphasis on ensuring that the telehealth system was accessible to the patient, no matter what their abilities were (Figure 13). This participant designed an app which provided multiple forms of feedback to the participant, including haptics, graphics, color-coding, audio and text-based information. The participant described that the app had to be accessible, and that some people who are color-blind would not be able to rely on color coding feedback, and that people with hearing loss would not be able to hear auditory feedback. In addition, this participant discussed the emotional and cognitive load that receiving medical information can have on a patient and discussed that the app would provide synthesized information to the patient—such as “good” instead of providing the raw data itself (such as a temperature or blood pressure numeric reading). The participant explained that sometimes patients are told a number by the medical professional, but that neither the child nor the parent(s) might understand what that number means, which can cause anxiety. To decrease patient anxiety and to increase clarity of information, it was important for the app to let the patient know if the sensor reading was something that they should be concerned about or not.

Another participant designed a system to address the problem of the patient being unable to communicate with their voice by designing a “brain reader,” in which the patient is able to communicate with their doctor by wearing a solar-powered hat which can read and speak their thoughts (Figure 14). This participant’s brain reader design and the hood thought reader design discussed earlier (Figure 4) emulate brain-computer interface (BCI) technologies, which are being developed to help people who are “locked in” and are unable to speak with their voice or use their hands to access mainstream technologies. Unlike most current BCI technologies, the brain reader hat design allows the child user to communicate either through voice output or through images. Here again, we see our participants generate creative design solutions which have the potential to be used by both adults and children.

Discussion

We framed our study to address the needs of a pediatric patient with a communication disability. Yet, many of the design solutions our participants created incorporate features that will, in fact, be useful to a broad population—including both pediatrics and adults, as well as both able-bodied and disabled individuals. In our discussion, we focus on some of the key lessons learned from our participants’ designs and how they can be practically incorporated into existing and near-future telehealth systems.
Telehealth technology should be negotiated at the family level, not dictated by design

Our participants naturally designed technologies in which the user is at the center of the design. The use of the technologies and the information shared are controlled by the user. In the case of a pediatric patient, the user could be defined as the child themself and/or their parents. Regardless of whether the child or their parent are the primary user, all designs were oriented towards being child-friendly, mirroring the experience of in-person pediatrician visits. It is important to note that our participants did not impose their views on who should ultimately have control of the technology, the child or parent(s), but left their designs flexible so that the choice could be made at the family level. We recognize that research has investigated how and when adolescents should have access to their health data and that opinions and policies on the topic vary. In the case of telehealth technology design, our participants assumed that the choice of access would be up to the families themselves, and not dictated by the technology designers. We encourage telehealth designers to adopt a similar design philosophy, in which the pediatric telehealth system can be utilized by both adults and children. We suggest that using this design approach will be useful not only to families, but also to caregivers of adults, and to individual adult patients.

Telehealth technologies should provide a holistic view of the patient’s current health (using sensor and visual input)

Our participants’ telehealth designs focus on the importance of obtaining the necessary information a physician would need in order to provide medical care to the patient. Participants created a variety of systems to capture medical data typically taken during an office visit, incorporating familiar tools such as thermometers. In some cases, such as with the drone design and the virtual reality robot, the physician was able to see and communicate with patients in real-time and could help assist in capturing the medical data needed. Surprisingly, none of our participants incorporated existing wearables and sensors, such as an AppleWatch or Fitbit, into their designs. However, the app designed by one participant and the removable sensor designed by another participant essentially function in a similar way to existing wearables, with the difference being that the participants’ solutions are designed specifically for the purposes of capturing data for telehealth visits. We recognize the use of mainstream wearables to gather medical data is an emerging area of exploration. Some healthcare providers have used home-based devices, such as blood-pressure cuffs and Fitbit data, during COVID-19 in an attempt to provide more holistic care through telehealth. Our findings for the future of telehealth suggest that healthcare providers can assist patients in the correct usage of a variety of sensor technologies as tools to supplement telehealth visits. Our participant’s YouTube design example is one way in which educational materials regarding the use of home-based medical devices could be shared in a pediatric-friendly way. We encourage telehealth designers to consider how they might incorporate educational information into the proper use of home-based devices and wearables so that the data collected could be useful for telehealth visits. The educational information would also need to include explicit instructions on how users choose to share their home health data with medical staff. As our participants’ designs indicate, the choice of sharing health information collected at home should ultimately be up to the family.

In addition to health data, another common design theme is the ability for the physician to view more than just the face of the patient during the telehealth visit. Our participants’ designs included views of the patient that are directed by the patient (e.g. the handheld video device with x-ray mode) and video solutions directed by the physician (e.g. the drone and robot). Based on these early conceptual designs, we suggest that telehealth systems be designed with the option for both patient-controlled and physician-controlled views. The interactive 3D body design, for example, could easily incorporate both physician and patient controlled views. Once again, we can imagine scenarios in which having a comprehensive view of the patient would be useful beyond the scenario described in our design prompt. We anticipate that adults, children, able-bodied and disabled populations would utilize one or more of these tools during a telehealth visit with their doctor.

Telehealth technologies should be accessible using multiple communication methods

Our participants’ designs show how telehealth systems can incorporate multiple modalities of communication. Children do not communicate in the same way as adults and we do not think that telehealth systems should be designed with only one communication approach available. Garth et al.’s research showed that pediatricians deliberately communicate at the child’s level to foster the pediatrician-patient relationship during in-person office visits. Telehealth systems should afford for the same types of communication adaptations. Telehealth systems should capitalize on communication methods children prefer, such as using asynchronous videos with commenting for sharing information. We encourage designers to embrace a multi-modality communication approach into telehealth technologies. Designing for one communication method alone is not enough. Current video conferencing
platforms can be modified to more explicitly include multiple communication modalities, such as improving
chat/text-based communication by making this option more visually accessible, and by incorporating phrase and
word prediction and incorporating emojis. Emojis could be particularly useful in patient-provider communication to
add more descriptive information into how the patient is feeling, which can benefit children, adults, people who are
literate and people who have limited literacy. Emojis could also be used to supplement traditional pain scale
questions as part of the telehealth exam.

Our design scenario of a pediatric patient with a disability prompted our participants to creatively think about other
accessibility features that could be incorporated into telehealth design. Using multiple modalities such as color
coding, text, visuals, haptics, and auditory feedback within the telehealth system can benefit many different potential
users in a variety of situations.

**Limitations and Future Work**

Our study is an exploratory study and more work needs to be done before any of our findings can be put into practice.
We have a small participant size, which is appropriate for this type of early-stage study, but a larger participant size
would be needed for future stages of research, such as prototype design evaluations. We also recognize that while our
scenario featured a child with disabilities, we did not include any design participants with disabilities. We believe that
future work in this area, especially early-stage prototype designs, should include participants with disabilities.

**Conclusion**

We conducted co-design sessions with six children on the topic of telehealth. Using a design scenario of a child with
communication disabilities, we explored how telehealth technologies could be improved. We found that children’s
ideas for telehealth systems were both provocative and practical. Our participants considered the perspectives of the
entire family’s use of telehealth in their designs, creating designs which could be used by both adults and children.
Moreover, our participants created designs which emphasized privacy measures controlled by the family as part of
the health data sharing process. Our participants were empathetic towards Pat, the character in our design scenario,
and as a result they designed solutions which could accommodate different communication and physical abilities.

Our study is an exploratory study of the potential of telehealth in the near-future. We believe that telehealth systems
that do not embrace concepts such as user-driven personal health data will ultimately be left-behind as the use of
telehealth grows. Fiks et al. argued that COVID-19 has created an opportunity for defining the future of pediatric
telehealth⁵. Our six co-design participants have provided inspiration for what the future of telehealth could be: user-
driven, flexible, and accessible technologies.

**Acknowledgements**

We sincerely thank our co-design participants and their families.

**References**

2. Heylighen A, Rychtarikova M, Vermeir G. Designing spaces for every listener. Univers Access Inf Soc. 2010
   Computer Science).
   participation of people with disabilities in an era of telehealth. J Am Med Inform Assoc [Internet]. 2020 Nov
5. Fiks AG, Jenssen BP, Ray KN. A Defining Moment for Pediatric Primary Care Telehealth. JAMA Pediatr.
   2021 Jan 1;175(1):9.
   Mar 1;52(6):839–51.
   Reviewing and Communicating about Radiology Imaging Studies. Proc SIGCHI Conf Hum Factors Comput
8. Hong MK, Wilcox L, Machado D, Olson TA, Simoneaux SF. Care Partnerships: Toward Technology to


Robot takes contact-free measurements of patients’ vital signs [Internet]. MIT News | Massachusetts Institute of Technology. [cited 2021 Feb 6]. Available from: https://news.mit.edu/2020/spot-robot-vital-signs-0831


A Framework for Inferring Epidemiological Model Parameters using Bayesian Nonparametrics

Oliver E. Bent, PhD, Charles Wachira, BSc, Sekou L. Remy, PhD, William Ogallo, PhD, RPh, Aisha Walcott-Bryant, PhD
IBM Research Africa, Nairobi, Kenya

Abstract

The use of epidemiological models for decision-making has been prominent during the COVID-19 pandemic. Our work presents the application of nonparametric Bayesian techniques for inferring epidemiological model parameters based on available data sets published during the pandemic, towards enabling predictions under uncertainty during emerging pandemics. We present a methodology and framework that allows epidemiological model drivers to be integrated as input into the model calibration process. We demonstrate our methodology using the stringency index and mobility data for COVID-19 on an SEIRD compartmental model for selected US states. Our results directly compare the use of Bayesian nonparametrics for model predictions based on best parameter estimates with results of inference of parameter values across the US states. The proposed methodology provides a framework for What-If analysis and sequential decision-making methods for disease intervention planning and is demonstrated for COVID-19, while also applicable to other infectious disease models.

Introduction

The COVID-19 pandemic is one of the most important public health concerns of the twenty-first century. Globally, more than 118 million COVID-19 cases and 2.6 million COVID-19-related deaths had been reported by March 2021. The United States, with more than 29 million COVID-19 cases and 540 thousand deaths over the same period, disproportionately bears most of the COVID-19 burden. A key characteristic of COVID-19 is that multiple timely interventions along with behavioral compliance to the interventions by populations are critical to contain the spread of the disease. However, although the intervention options and behavioral aspects may be limited, tackling the COVID-19 pandemic has, in part, been hindered by the insufficient understanding of data and models that can adequately inform the decision-making process of stakeholders.

Overall, non-pharmaceutical interventions (NPIs) such as quarantine, mask-wearing, and school closures have played a critical role in substantially slowing the spread of COVID-19 across different populations and geolocations\(^\text{1}\). To this end, research efforts have been directed towards curating NPI datasets from publicly available data sources\(^\text{2,3}\), tracking government responses by scoring the stringency of imposed NPIs\(^\text{4}\), and developing a plethora of models for forecasting COVID-19 outcomes to support policymaking\(^\text{5-8}\). At the same time, researchers are actively investigating the dependence of COVID-19 on human behavior. For example, studies have investigated human mobility as a measure of the stringency of travel restrictions during COVID-19\(^\text{9}\) and as a predictor of the spread of COVID-19\(^\text{10}\). Unfortunately, however, little is known about the interactions between imposed NPIs, human behavior, and COVID-19 outcomes especially to improve the calibration and learning of model parameters in compartmental models.

The overarching goal of this study is to introduce a framework and method for inferring epidemiological model parameters by integrating non-traditional epidemiological model drivers and surrogates as inputs during the model calibration process. Specifically, we demonstrate how the stringency of NPIs and human mobility data can be used to train a Gaussian process regressor, and subsequently infer values of parameters like the transmission rate in an SEIRD (susceptible, exposed, infectious, recovered, dead) compartment model using example states in the United States. A key contribution of this work is the ability to conduct What-If and scenario analyses that decision-makers can use to estimate the transmission rates given based on NPI stringency and human mobility.

Related Work

Besides compartmental models such as the SEIRD, a wide range of modeling efforts has been directed towards understanding and controlling COVID-19. For example, pharmaceutical intervention discovery has employed crowdsourcing in pandemics, with the use of Machine learning and computational tools to rapidly triage compounds and design...
synthetic routes, reducing the time it takes to find treatments with the most promise\textsuperscript{11,12}. Machine learning-based models, trained on specific biomolecules, have also offered inexpensive and rapid implementation methods for the discovery of effective viral therapies\textsuperscript{13} along with direct applications to the design of clinical trials and experiments through the use of causal inference and Reinforcement learning methods\textsuperscript{14}. Recently, researchers have also developed Non-Pharmaceutical Intervention repositories which have been extended through automated extraction of global NPI details from Wikipedia articles using NLP\textsuperscript{2} which run in parallel to intensive manual data collection projects\textsuperscript{4}.

Early in the pandemic, Google and Apple released human mobility data sets that capture mobility patterns for specific types of locations or activities. Google’s COVID-19 Community Mobility reports\textsuperscript{15} contain data that represents the percentage difference in mobility from a "pre-pandemic" baseline for six types of locations: retail and recreation, supermarket and pharmacy, parks, public transport, workplaces, and residential. Apple\textsuperscript{16} offers mobility data based on the number of requests for directions for three types of modes of transport: driving, walking, and public transport. Such data are critical for the pandemic and serve as a proxy measure in understanding the public behavior in response to the pandemic and adherence to NPIs imposed by governments. Several researchers have leveraged mobility data and other environmental data to support modeling efforts and infer public behavior throughout the pandemic. For example, in Nouvellet et al.\textsuperscript{17}, the relationship between mobility and COVID-19 transmission is explored, specifically the reproduction number using EpiEstem\textsuperscript{18}, for 52 countries at different phases of lockdown. In another study by Oliveira et al.\textsuperscript{19}, the authors explore the relationship between mobility and lockdown phases at different points during the pandemic, and describe disparities in this relationship across South American countries. Mobility data has also been used to explore its relationship to COVID-19 outcomes at a county-level or set of counties in Arizona by Wang et al.\textsuperscript{20}. Work by Kuo et al.\textsuperscript{21}, combines mobility data with demographic and environmental data into a hybrid ML model to predict incidence at different response phases (e.g., lockdowns) during the pandemic. While these methods incorporate mobility data into understanding the pandemic across various geographies, they do not provide an inference framework that includes the stringency of NPIs and mobility to infer parameters for compartmental models, and further are not directly used in What-If analysis and intervention planning.

Long-short term memory (LSTM) networks are increasingly popular as a deep learning alternative to traditional epidemiological modeling. LSTMs are a type of recurrent neural networks that can learn long-term temporal dependencies\textsuperscript{22}. They are particularly popular because of their ability to address the limitations of traditional time series forecasting techniques such as the need for restrictive assumptions of linearity in data, resulting in more accurate models. These techniques have been used by researchers for modeling COVID-19 time series data. For example, Chimmula and Zhang\textsuperscript{23} use LSTMs to forecast future COVID-19 cases and possible stopping time in Canada, and the rest of the world. Similar approaches have also been applied by other researchers\textsuperscript{24–28}. Kirbas et al.\textsuperscript{7} and Devaraj et al.\textsuperscript{29} found LSTMs to be more accurate than techniques such as Auto-Regressive Integrated Moving Average (ARIMA) and Nonlinear Autoregression Neural Network (NARNN) in predicting COVID-19 cases. Among different variants of LSTM models for COVID-19 forecasting, researchers report better accuracy for bidirectional-LSTM models\textsuperscript{8,30,31}. Researchers have also combined LSTMs with other techniques\textsuperscript{32}. It is worth noting, however, that most of the techniques discussed above are typically black-box models that are often not readily interpretable or grounded in epidemiological processes.

Epidemiological Model Structures

The major forms of mathematical models for epidemiological systems may be classified into three groups:

- **Black-box** models are entirely data-driven models and their structure is not guided by any underlying physical mechanism. Such model types are common in machine learning and statistics. In either case, the model parameters are learned from observation.

- **White-box** models are considered to be mechanistic and attempt to describe the underlying physical mechanisms. As an example of white-box models, agent-based models are constructed based on interactions between multiple agents of the epidemiological system and aim to create a closer resemblance to real-world dynamics.

- **Grey-box** models consist of mixtures of phenomenological (black-box) and mechanistic (white-box) models.
For our work, we will consider all models as black-boxes, epidemiological models therefore acting as a set of exposed parameters, which allows us to switch seamlessly between the use of different model types and between diseases. This ground-up approach of extensibility and scale is perhaps rooted in the premise that much of machine learning considers models as black-boxes across application domains. In the case of epidemiology, challenges in reproducibility, and lack of sharing or a set of standard benchmarks, limit the amount of work that effectively compares the results of different modeling approaches. Consequently, instead of focusing on the analysis of epidemiological model structures, we look at mappings between inputs and outputs from various models for forecasts or predictions under uncertainty, ultimately to be exploited for the task of decision making.

While the internal mechanisms of the model used remain as black-boxes, different approaches return deterministic or stochastic results. This is of note with regards to how uncertainty may be handled from the model. A deterministic epidemiological model principally takes the compartmental form of Kermack and McKendrick. Such models are deterministic in the sense that, once we know the population numbers of each compartment (initial conditions) and the rate constants between the compartments, the system is fully determined and has a single outcome or prediction, while for any real-world scenario, reducing the problem in such a way is unlikely to be realistic, or have strong predictive performance. We may inject some randomness into deterministic models through quantifying uncertainty in the model parameters or initial conditions. Deterministic compartmental models tend to be suited for large well-mixed populations, with relatively small numbers of recovered individuals. Typically a deterministic approach may be evaluated much faster, as differential equations have efficient solvers with a single outcome being computed. Stochastic simulations may need many more computational cycles to model the atomic transitions of a stochastic agent-based model in comparison.

Taking a common footing for both modeling approaches through uncertainty quantification can lead us to the following classifications. Uncertainty in predictions arising from the uncertainty in the model parameters is called epistemic uncertainty. This could lead to a distribution of outputs from a deterministic model if parameter distributions were defined on the model parameters, or based on a stochastic model’s own description of epistemic uncertainty. This kind of uncertainty may be better quantified with more data or observations. Consequently, epistemic uncertainty should be higher in regions of little or no training data and lower in regions of more training data. Uncertainty arising from the inherent noise (or imprecision) in the disease dynamics is an example of aleatoric uncertainty. It cannot be reduced if we get more observations; in principle, we have to accept that the future is uncertain.

**Gaussian Process Regression**

The Gaussian Process (GP) is a well-known statistical learning model, used extensively for probabilistic non-linear regression. The task of regression aligns with parametric, phenomenological, or black-box epidemiological model design for prediction, specifically how predictions as forecasts of future epidemiological state may be learned. Although Gaussian process regression (GPR) is not a standard modeling approach in epidemiology, there are examples of its use in geospatial modeling. We focus on motivating its use in a pipeline for temporal decision-making using epidemiological transmission models. GPR targets a functional approximation for an underlying stochastic process, a natural fit for capturing uncertainty from epidemiological model training examples.

The outputs used from epidemiological simulations are considered as stochastic, either through mechanisms of stochasticity within the model or based on stochastic distributions of model input parameters, samples of which generate a stochastic output. Firstly a general overview of GPR is necessary before generating the appropriate mapping for the learning of model parameters for prediction tasks.

A GP is fully specified by its mean function \( m(x) \) and covariance function \( k(x, x') \). This is a natural generalization of the Gaussian distribution, whose mean and covariance are a vector and matrix, respectively. The GP entertains a probability distribution over some (real) underlying stochastic process \( f(x) \). Defining:

\[
    m(x) = \mathbb{E}[f(x)]
\]

\[
    k(x, x') = \mathbb{E}[(f(x) - m(x))(f(x') - m(x'))]
\]
K runs. The covariance pairings of these random variables represent the value of the stochastic process \( f \) at a location \( x \in X \). The location variable in our application is over epidemiological model parameters. While considered a Bayesian nonparametric learning technique, GPR does involve the specification of a limited set of parameters, almost entirely contained with the covariance kernel function. Judicious choices of kernel function and parameter values need to be made for the successful application of GPR.

**Inference of Predictions based on Changing Epidemiological Model Parameters**

The specification of the covariance function implies a distribution over functions. To see this, we can draw samples from the distribution of functions evaluated at any number of points; in detail, we choose a number of input points denoted \( X_* \).

\[
f_* \sim \mathcal{N}(0, K(X_*, X_*)^{-1})
\]

In which \( X_* \) is a vector of our test inputs (or locations) of the regression, \( f_* \). Therefore, without observations or training inputs, this fully specifies our Gaussian process prior.

We consider an underlying stochastic or noisy epidemiological function \( y = f(x) + \epsilon \), \( \epsilon \) being additive independently identically distributed (i.i.d) Gaussian noise with variance \( \sigma^2 \), Here we’ve assumed that the underlying stochastic process is unbiased and \( m(x) = 0 \). Without other reasonable prior knowledge e.g. periodicity this is a reasonable assumption, along with our beliefs of the underlying process that we have specified through the GP covariance function. Our stochastic epidemiological function \( y \) is therefore distributed such that:

\[
y \sim \mathcal{N}(0, K(X, X) + \sigma^2 I)
\]

This now allows us to specify the joint distribution of the observed training values \( X \) and the function values at the test locations \( f_* \) under the GP prior.

\[
\begin{bmatrix} y \\ f_* \end{bmatrix} \sim \mathcal{N} \left( 0, \begin{bmatrix} K(X, X) + \sigma^2 I & K(X, X_* ) \\ K(X_*, X) & K(X_*, X_*) \end{bmatrix} \right)
\]

(3)

If there are \( n \) training points and \( n_* \) test points then \( K(X, X_*) \) denotes the matrix of the covariance function evaluations at all pairs of training and test points and \( X \) the vector of our training inputs or results from epidemiological runs. The covariance pairings of \( K(X, X), K(X_*, X) \) and \( K(X_*, X) \) follow respectively from test points \( (X_*) \) and training points \( (X) \).

Through constructing the conditional distribution based on the assumption of Gaussian prior observations this gives us the following:

\[
f_* | X, y, X_* \sim \mathcal{GP}(m(x_*), k(x_*, x'_*))
\]

In the case of a zero mean function prior this reduces the prediction functions to:

\[
m(x_*) = K(X_*, X)[K(X, X) + \sigma^2 I]^{-1}y
\]

(4)

\[
k(x_*, x'_*) = K(X_*, X_*) - K(X_*, X)[K(X, X) + \sigma^2 I]^{-1}K(X, X_*)
\]

(5)

The above results have direct solutions conditioned on our GP hyper-parameters \( l, \sigma^2 \) and \( \sigma^2 \), which we shall term \( \theta_{GP} \). While judicious approximations can be made for these values, we rely on an optimization scheme to fit them based on observation across tasks. This may be done through maximization of the log marginal likelihood:

\[
\log p(y | X, \theta_{GP}) = -\frac{1}{2} y^T (K(X, X) + \sigma^2 I)^{-1} y - \frac{1}{2} \log |K(X, X) + \sigma^2 I| - \frac{n}{2} \log(2\pi)
\]

(6)
Performing Maximum Likelihood Estimation (MLE) based on maximization through gradient-based optimization, recovers an empirical Bayes estimate to $\theta_{GP}$. While full Bayesian inference may be performed to generate the posterior parameter distribution, the computations are analytically intractable, approximate methods are therefore deployed, as exemplified by Markov Chain Monte Carlo (MCMC) algorithms for sampling the GP prior.

**Learning Model Parameters from Observation**

A calibrated model implies that the model parameters ($\theta_m$) have been selected, either as single values or distributions. This calibration may have been performed by the model developer for the specific disease and location, for which the model has been specified. The process of calibration takes the designed model structure and exposes meaningful parameters ($\theta_m$) to tuning such that a model achieves higher predictive performance. A model developer may have designed a model with physical parameters in mind and reference the literature for these parameter values, and simply look for the model structure to represent observations. Alternatively, a grid-search of parameter values may be performed, which exhaustively evaluates model runs and fixed interval perturbations of the parameter values, specifying a hyper-dimensional grid in the number of parameters. The vertex associated with the parameter set that provides a best-fit (or minimum loss) then being chosen. In this work, we advocate for an algorithmic approach to achieve best-fit, which may both improve the efficiency of our search for a parameter set, under computational constraints, while also providing additional information with regards to parameter sensitivity and allows us to incorporate additional information around parameter distributions.

We assume a predefined model structure with exposed model parameters ($\theta_m$), and consider the mechanisms of the model as hidden under the black-box abstraction. Simply being given access to a disease model structure almost always requires tuning or calibration of that model to generate meaningful predictions. Therefore, we develop mechanisms that can efficiently achieve this, enabling large-scale use and comparison of epidemiological model predictions.

The size of the parameter set being calibrated for a mechanistic epidemiological model is often in the order of $10^3$. When performing calibrations for models of the same disease and same location, a mapping of any common parameter sets and to get similar model performance is noted to be a complex task. In addition to this task, we consider the task of separately calibrating single models for a new geography. To perform such calibrations, reference observational data such as cases, deaths, and vector numbers are required. We may then interrogate further the impact of parameter values with regards to this observational data and model outputs.

**Bayesian optimization**

Bayesian optimization is a principal global optimization method driven by the exploration-exploitation paradigm and used in black-box optimization of computationally expensive functions. Priors on our parameters may build in human expert knowledge from epidemiologists; and an acquisition function which guides the exploration process, reducing unnecessary evaluations of expensive objective functions while maximizing information. Importantly, Bayesian optimization leverages approximations in the form of a surrogate model or response surface to remove the computational expense of direct evaluations of the objective function. This surrogate model is almost always a Bayesian nonparametric model, the Gaussian Process. While other surrogate modeling methods exist for sequential model-based optimization, we have developed a consistent set of approaches based on the GP posterior and will not divert, though methods such as Parzen Estimator approaches exist. As epidemiological models tend not to be written with closed-form solutions, we are led to approaches that treat the environment against which a candidate solution is being evaluated as a ‘black-box’. Specifically we seek to approximate an optimal solution given that the:

- Objective function ($f(x)$) is unknown, non-linear, non-convex, non-derivative and stochastic.
- Candidate solutions ($x$) are a high dimensional vector from the set of possible actions. ($A \subset \mathbb{R}^d$), typically with $d \leq 20$.
- Evaluation of $f(x)$ is computationally very expensive.
Therefore by framing the process of exploring model parameter solutions $x$ as an optimization problem:

$$\max_{x \in A \subset \mathbb{R}^d} f(x)$$

(7)

For the purposes of calibration, these candidate solutions $x$ exist across dimensions $d$ (often a rectangular hyperspace with uniform ranges $x \in \mathbb{R}^d : a_i \leq x_i \leq b_i$), which are evaluated towards the maximization of some objective function $f(x)$. The objective function returns a scalar reward that quantifies the performance of the solution.

In outline, we use Bayesian optimization as a sequential optimization schedule that combines our GP surrogate model with the acquisition function which is detailed in the following section.

**Algorithm 1:** Bayesian optimization Steps

```
for $t = 1, 2, ..., n$ do
    select new $x_{t+1}$ under acquisition function $\alpha$;
    $x_{t+1} = \arg\max_{x \in A} \alpha(x; D_t)$;
    run model simulation to provide $y_{t+1}$;
    new observation $D_{t+1} = \{D_t, (x_{t+1}, y_{t+1})\}$;
    Update GP Posterior: mean and covariance
end
```

In Algorithm 1 we have introduced the notion of an acquisition function ($\alpha$), this may have several functional forms. Popular acquisition functions are UCB, EI and Thompson Sampling - they may be grouped in the optimistic, improvement-based and information-based search methods respectively

**Loss Function**

A loss function ($L$) maps our parameter estimates to a scalar value ($L : \mathbb{R}^d \rightarrow \mathbb{R}$) from function outputs against which Bayesian optimization may be performed. In this framing of a loss function for calibration, the optimization minimizes the loss. The search for a minimum Normalized Root Mean Square Error (NRMSE) is equivalent to the maximum likelihood estimate, a point estimate of calibration parameters.

$$L = NRMSE = \sqrt{\frac{1}{n} \sum_{t=1}^{n} (y_t - \hat{y}_t)^2}$$

(8)

Such a loss function is standard with epidemiological model calibration and is not an innovation but included for completeness as to which we demonstrate results of a Bayesian optimization process.

**Results using a SEIRD Compartmental Epidemiological Model**

We used stringency index and mobility data to generate an input time series data set for all the states of the United States of America. The stringency index quantifies the strictness of government-imposed NPIs in a location while the mobility data captures the daily changes in movement compared to pre-pandemic baselines. For each state, the study period ranged from the day it reported its first case to August 3, 2020, to capture the first wave of the epidemic. We used an SEIRD model and learned parameter values for the defined periods. These parameters included: death rate $\delta$, recovery rate $\gamma$, and the transmission rate $\beta$. The parameters were learned by applying Bayesian Optimization to fitting the model to the recorded confirmed cases and deaths. We also used six values of $\beta$, i.e. five changes in the transmission rate during the calibration duration. We ran 20 calibration events for all the states and used the Mean Absolute Percentage Error (MAPE) of 4 weeks prediction of confirmed cases and deaths to identify 47 states where the MAPE was below 4. We trained a GP to estimating the impact of stringency and mobility on $\beta$.

Figure 1 shows the predicted and observed confirmed cases, as well as the learned $\beta$ and $R_0$ values for the selected states of Florida, Georgia, and New York. The projected values generated from the calibrated parameters closely follow

222
the observed data. There are some similarities and differences to note between these states. Each state is observed to have a high (>1) transmission rate at the beginning of the pandemic that starts to reduce later in the pandemic. New York was able to get $\beta$ down to very manageable levels in May, but like Florida and Georgia, by the end of July, values were increasing once more. After their first reported cases, Georgia and Florida followed similar trends, both approaching 100k cases in July, even though the trajectories for $\beta$ follow different paths. It is clear that the dynamics of the pandemic in each of the states have followed distinct courses, and both the governance and the practices of the population are certain to have played critical roles. This is why we selected the stringency and mobility values as features in our inference method, as they serve as a proxy to how governments and the population have responded.

**Learning Parameter Distributions from Data**

In Figure 2 we present the one- and two-dimensional functions learned from the data for the State of New York. Figures 2a and 2b present the mean and covariance for the GP trained to estimate the relationship between $\beta$ and stringency and mobility respectively. Of note in Figure 2a is that in New York, stringency values have a narrow range between 0.5 to 0.8, implying that the state remained fairly strict throughout the study period. As would be expected, there seems to be a negative correlation between the stringency and the predicted $\beta$ values. Figure 2b suggests that there is minimal correlation between observed mobility and predicted $\beta$ values. This implies that if there is a relationship between the two, that relationship is either complex or confounded by other factors that warrant further investigation. These include the timing of the enforcement of stringency measures that affect mobility, the context of which the measures were implemented, the measures that were already implemented, and the number of cases that were prevailing. All of these influence the assessment of the impact on transmission rate, making causal inferences difficult to extrapolate.

Figure 2c presents the mean of the surface learned from mapping both of these values to $\beta$. As with the other two subfigures, there is a strong signal indicating a relationship between these values, and the interaction between stringency and observed mobility with $\beta$. The mean value of $\beta$ represented in the pixel value in the range $[0, 1]$. 

**Figure 1:** Calibrated SEIRD model output of the total confirmed cases as well as the learned $\beta$ and $R_0 = \frac{\beta}{\gamma}$ values.

**Figure 2:** Plots of the mean and covariance (including the parameter samples) learned from GP regression for the State of New York. Figure 2a illustrates the relationship between stringency values and calibrated transmission rates ($\beta$). Figure 2b illustrates relationship between observed mobility and $\beta$. Figure 2c illustrates the relationship between both stringency and observed mobility with $\beta$. The mean value of $\beta$ represented in the pixel value in the range $[0, 1]$. 

and mobility is also at play. In each of these cases, one might consider these figures as a prescriptive yet noisy way to infer the impact of changes to the way that governments or populations respond to the pandemic and the ability of the disease to propagate through the state. If the learned signal is truly representative, it provides a means to perform What-If scenario evaluations, permitting decisions to be made based on the expected impact on transmission. Specifically, the regressed GP can be used to define the values of input parameters (in this case $\beta$, but one can readily see that others could have been learned as well). Here $\beta$ would be a distribution defined by the selected stringency and/or observed mobility.

Even though different states have different approaches to managing the pandemic, we can estimate the impacts of their strategies. Aggregating the samples from multiple states provides one mechanism to increase the range of predictions and to incorporate additional information for regions of the surface which are captured in any given states. This aggregation might increase the uncertainty in the projected impact in some regions, which is important as it represents a diversity of outcomes that are possible depending on the state, the pandemic context, and how their population might respond to interventions. To aggregate, we train the GP with samples from 47 states which have Mean Absolute Percentage Error (MAPE) of less than 4 on 20 independent calibrations events of the SEIRD model. Using the resulting GP, predictions for the values of $\beta$ were generated and used as input for the compartmental model.

Performing Model Predictions with Learned Parameter Distributions

The learned parameter distribution was used in the SEIRD model to predict confirmed cases for the states of Florida, Georgia, and New York. As shown in Figure 3, a four-week prediction from August 3, 2020 is plotted for SEIRD-model, where the last model calibrated beta value is held constant for the prediction period; and for GP-regression, where $\beta$ values are learned from the GP. The observed confirmed cases are also illustrated. The GP-regression plot despite over-predicting compared to SEIRD-model plot, carries information about how strict the interventions are and how people are moving as a factor of the transmission parameter, which is key in intervention planning.

Predictions were performed for one, two, three, and four weeks from the last calibration date. For each prediction, the MAPE was computed on the predicted confirmed cases and deaths. Table 1 shows the MAPE for the 3 states and each of the experiments.

Table 1: Mean Absolute Percentage Error (MAPE) measuring the accuracy of SEIRD model ($\beta$ values learned directly from calibration) and GP-regression ($\beta$ values influenced by stringency and observed mobility) in forecasting COVID-19 cases within 4 time periods, and across 3 states

<table>
<thead>
<tr>
<th>State</th>
<th>Experiment</th>
<th>Week 1</th>
<th>Week 2</th>
<th>Week 3</th>
<th>Week 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Florida</td>
<td>SEIRD-model</td>
<td>0.057909</td>
<td>0.130846</td>
<td>0.208969</td>
<td>0.290546</td>
</tr>
<tr>
<td></td>
<td>GP-regression</td>
<td>0.066345</td>
<td>0.169851</td>
<td>0.295689</td>
<td>0.434123</td>
</tr>
<tr>
<td>Georgia</td>
<td>SEIRD-model</td>
<td>1.076135</td>
<td>1.09348</td>
<td>1.10957</td>
<td>1.117028</td>
</tr>
<tr>
<td></td>
<td>GP-regression</td>
<td>1.081247</td>
<td>1.1595</td>
<td>1.33907</td>
<td>1.33907</td>
</tr>
<tr>
<td>New York</td>
<td>SEIRD-model</td>
<td>0.328833</td>
<td>0.363539</td>
<td>0.39194</td>
<td>0.433136</td>
</tr>
<tr>
<td></td>
<td>GP-regression</td>
<td>0.328997</td>
<td>0.364593</td>
<td>0.395165</td>
<td>0.437741</td>
</tr>
</tbody>
</table>
Conclusions

In this work, we have presented an application of Bayesian nonparametrics in the form of the GP to provide a surrogate model in calibrating epidemiological model parameters via Bayesian optimization. Further, we presented a process to tie parameter calibration to relevant model drivers, Stringency and Mobility, thereby capturing changes in the model parameters. These changes may themselves be inferred through a GP regression against relevant available data. Our results demonstrate that inferred model parameter values of transmission rate ($\beta$) for US states approach the predictive performance of the best-fit $\beta$ parameter values for each state. Finally, the steps presented in this work are provided as open-source tools for the reader to interact with and extend the presented Bayesian nonparametric approaches to their own model structures and input datasets. Future work can focus on hyperlocal modeling of extended pandemic periods with different epidemic waves across multiple geographical locations.

References


[18] MRC Centre hosted within the Department of Infectious Disease Epidemiology at Imperial College London. A tool to estimate time varying instantaneous reproduction number during epidemics. https://github.com/mrc-ide/EpiEstim, 2013.


Comparing Older and Younger Adults Perceptions of Voice and Text-based Search for Consumer Health Information Tasks

Karen Bonilla, BS¹, Brian Gaitan, BS¹, Jamie Sanders, BS¹, Noami Khenglawt, MS¹, Aqueasha Martin-Hammond, PhD¹

¹Indiana University – Purdue University Indianapolis, Indianapolis, IN, USA

Abstract

The increased prevalence of voice search presents opportunities to address consumer challenges accessing online health information. However, it is essential to understand how users’ perceptions of voice affect their search processes for health information, concerns, and different scenarios for using voice for health information tasks. We conducted semi-structured interviews with 16 younger (18-25) and older (60-64) adult participants to understand and compare their perceptions of using voice and text-based search for non-health-related and health-related tasks. While most participants preferred traditional text search, younger adults were not inclined to use voice search for health information due to concerns about privacy, credibility, and perceived efficiency in filtering results. Older adults found voice search potentially beneficial for reducing manual query generation burdens; however, some were unsure of how to use the technology effectively. We provide a set of considerations to address concerns about voice search for health information tasks in the future.

Introduction

While consumer-facing voice technologies (e.g., browser-based voice search, mobile assistants) have been around for some time, advancement in speech recognition and natural language understanding have led to a growth in the use of these technologies in everyday life. Some online voice-search technologies such as Google Voice and Microsoft Cortana allow users to navigate visual user interfaces by performing traditional keyword searches using their voice. Others such as Apple Siri and Amazon Alexa act as assistants to find information on behalf of the user. Each approach has pros and cons, but for each approach voice is perceived as a more natural and convenient way of interacting with information and for completing search tasks.

The availability of voice technologies and their perceived efficiency has also led to an increased interest in using these technologies to support health information tasks. Several researchers have investigated the potential usefulness of voice to help different users obtain information and resources regarding their health. Yet, early findings are mixed as some users see potential while others feel that using voice for health tasks poses safety concerns. Despite concerns, research in this area is growing leading to a need for closer examination and additional empirical evidence of users’ experiences with voice for health search tasks to identify ideal use cases and ways forward.

In this paper, we build on prior work to examine older and younger adults’ perceptions of voice search for consumer health information tasks. We purposefully recruited to compare older and younger adult experiences due to potentially different generational and technology experiences. We conducted semi-structured interviews with 16 participants (6 younger adults and 10 older adults) that searched for online health information using text-based input and voice. For the purposes of this study, we did not limit voice search to a particular type of device (e.g., online voice search vs. voice assistant) because we wanted to understand a wide range of experiences with the two different online voice search methods. We found that both older and younger adults utilized the Internet as a supplement to doctor’s advice and explore alternative treatments. Both appreciated the convenience of being able to search for health information online and felt that it provided them an opportunity to become a more informed patient. However, both groups experienced challenges with searching for information online particularly with gauging the credibility of the information and health literacy. When it came to voice however, younger adult participants were more critical, sharing that they believed that using voice did not provide the depth of insight needed to be able to compare information when searching for health topics. In comparison, older adults were more receptive and hopeful about using voice for health, suggesting that if some open issues could be fixed, they felt it would be more convenient than traditional text-based keyword search. We discuss our findings in relation to prior work and make suggestions for improving older and younger adults’ experiences with voice search for health in the future.
Related Work

The Internet has become a key resource for finding and accessing information and helping patients become more engaged in their care. However, for some tasks such as for health information search, finding and navigating information can be challenging. Over the years, there have been many efforts to improve users’ interactions with online health information through computational means, many of which have focused on older adults. For example, some researchers have explored guided search to assist older adults with navigating health information online. Other interventions have focused on improving older adult’s online health literacy. However, in recent years, there has also been widespread interest in examining how voice technologies can support health and wellness.

A review of the use of conversational agents in healthcare reveals that voice technologies and other conversational user interfaces are increasingly being explored to support healthcare tasks ranging from mental health to chronic disease management. With the increase in voice-based technologies available to consumers, interest in exploring technologies such as voice assistants for both clinical and home healthcare delivery is also growing. Studies have explored using emerging voice-technologies to support independence and wellness in the home of older adults. Several researchers have also explored voice technologies to support aging in place while others have examined the potential of voice technologies for navigating and providing access to different types of health information. For example, recently researchers have explored the use of voice assistant technologies for supporting healthcare delivery tasks during the pandemic noting the benefits and barriers of existing voice assistant devices.

For older adults and other users, voice-based technologies are often perceived as more efficient and are thought to provide advantages for supporting access to information among older adults and people with disabilities. Because of these perceived advantages, voice technologies have gained renewed attention in recent years to examine uses for informational tasks such as health search. For example, O’Brien and colleagues examined reviews of popular commercial voice assistant technologies to understand how older adults are using devices, and beyond typical uses such as entertainment and smart home control, they also found that older adults were using voice assistants for memory aids, companionship, and emergency communication. Nullam and colleagues found that low-income older adults perceived voice assistants as a potentially efficient way to improve their access to health information and resources at home but like others, had concerns about the privacy and security of health information shared with the device. On the other hand, early research suggests that using certain voice-based technologies might leave users susceptible to other risks. Bickmore and colleagues for example, found that while most commercially available voice assistant devices can make the search for health information more efficient, they also place users at risk for unsafe behaviors, particularly if the user acts on the advice without proper knowledge of how to use the information returned. However, despite concerns, some suggest that with improvement voice and other conversational technologies could transform how older adults engage with health information and resources. But, there is still limited understanding of the usefulness of voice search for health. In this paper, we examine younger and older adults’ experiences searching for health information online to compare their experiences and perceptions of using voice search for consumer health information tasks.

Method

We conducted semi-structured interviews with 10 older adults and 6 younger adults to understand their experiences using traditional, keyboard and emerging, voice search technologies for health-related and non-health-related tasks. In this section we discuss our purposeful recruitment method and our data collection and analysis procedures.

Participants

Participants’ ages ranged between 18 to 64. Because one goal of our study was to compare experiences, we purposefully recruited participants in the age range of 18-35 (younger adults) and above the age of 60 (older adults). Our reasoning for selecting these groups was to understand if differences in responses might occur based on generational experiences with technology. Older adult participants were recruited from a local senior center, while younger adult participants were recruited from the local community. Prior to their selection, potential participants were informed of the main inclusion criteria: experience using traditional, text-based keyboard or voice search and experience searching for health information online. Interviews were conducted either in person or remotely at a time of convenience for both participants and researchers. Interviews took place before the COVID-19 pandemic. Of the sixteen participants, nine of the older adult participants were female and one was male. Half had a college degree, either an associate’s or bachelor’s degree, two participants had a high school degree, one had vocational training, and the remaining two participants had college credit but no degree. Three older adult participants had physical impairments and two had visual impairments; however, these impairments did not prevent them from searching online.
using typed search and voice search. Of the ten older adult participants, seven used computers, four used mobile devices, and one used a tablet. For younger adults, six used a computer and four used mobile devices. Of the six younger participants, four were female and two were male. Half of the younger adult participants had bachelor’s degrees, two had a high school diploma, and one graduated with an associate degree. No younger adult participants reported a disability or impairment.

Figure 1. Participants preferred online search methods and experiences with voice search tools.

Older adults and younger adults preferred text-based keyword search for finding information online, however many older adults preferred to use different methods for different situations (See Figure 1). Younger adult participants used Google voice search or Siri while older adults used a wide spectrum of voice search tools including Google voice search, Microsoft Cortana, Amazon Alexa, and Siri. Google voice search was the most used voice search tool among both groups (See Figure 1).

Data Collection

Before commencing interviews, we obtained approval from the Institutional Review Board at Indiana University. During interviews, after informing participants about the study and their rights, we asked participants to complete a background questionnaire to collect demographic data and information about their current technology use for searching information online. We then engaged participants in semi-structured interviews. In interviews, we asked participants to reflect on their experiences by sharing particular instances where they searched for health and non-health related information using each approach. For each scenario they shared, we asked them to reflect on the benefits and challenges of the approach and what they might do differently, if the content (health vs. non-health related) or approach (voice vs. traditional) were changed. To fully understand the common deterrents and benefits of each approach (i.e., voice or text-based search) we asked participants to share their general Internet search practices and preferred approach (voice or traditional search), reason and processes for searching for health and wellness information, experiences with voice-based search, and experiences using voice search for health and wellness. To conclude the interview, we focused discussions on the effectiveness and relevance of information provided by their preferred search approach, the benefits and challenges of their preferred approach, and suggestions for future improvements. All interviews were recorded and then transcribed. Participants were asked for consent prior to the recording of the interview. Participants were provided with a $20 grocery gift card for their time.
Data Analysis

We used thematic analysis\(^{41,42}\) to identify emerging themes from the data. Our analysis process entailed two researchers independently noting common codes based on an initial review of each of the transcripts. Following the initial independent coding, the two researchers met to discuss and consolidate codes which led to an initial set of codes that were used in a second round of analysis to code data. Example codes included themes such as search preferences, device reasoning, challenges, and reasons for online health search. In addition to thematic codes, we also grouped the responses of older and younger participants to examine similarities and differences in their responses and approaches. The final stage of analysis included several meetings where the two researchers met to consolidate similar codes under high-level themes which correspond to the subcategories listed in the findings section.

Findings

We discuss and compare younger and older adults’ experiences with using voice and traditional, text-based search for health and non-health related tasks.

Experiences with voice and traditional, text-based search for everyday search tasks

While the focus of our interview was to investigate online health search within a broad spectrum, we also sought to compare participants’ experiences with the efficiency of voice search versus text-based search for different tasks. Both younger and older adult participants shared different experiences. During interviews, we asked participants about their preferences when searching online, not limited to health information tasks. Less than half of the older adult participants (N = 4) reported feeling comfortable using voice search and preferred it in comparison to traditional, text-based search. The other half of older adult participants were not opposed to voice search but preferred traditional search because they were more familiar with searching for information that way. Three of the older adult participants in this group mentioned they would be willing to use voice search more often if they had better instructions on how to use it. Some older adult participants therefore were hesitant to use voice technologies due to what they perceived as a barrier of entry, and they were not inclined to search online for instructions for a product they used infrequently.

All younger adult participants (N = 6) preferred traditional text search, with only two mentioning consistent use of voice search. Younger participants’ inclination towards traditional search was due to their belief that traditional text search was more efficient and accurate. In comparison to voice search, younger adult participants felt that when using traditional text search, there was less of a chance of their query being incorrectly interpreted. Participants also liked being able to manually type in their queries; any mistakes they made could be quickly fixed opposed to the potential additional steps to fix issues related to speech recognition errors. Participant 12, stated, “I think usually people want information as quickly as possible so if they have to do it manually by typing it out, they’ll do that. I don’t see the huge need for voice-based search engines unless someone has a disability, that’s where I think they’re necessary”. Security and reliability were also raised as concerns about using voice search. Participant 9 mentioned, “Because it [voice search] doesn’t feel totally reliable for me these days. I’m pretty sure that it’ll change in 15, 20 years when they develop and improve the technology, but I don’t like the way it works right now”. This participant’s criticism with voice search related to the prior concerns about being able to control the search process, participants felt that voice search provided less control for generating queries but also manipulating queries to customize the search. Participants felt that currently the lengths they needed to go through to correct a query mistake ended up outweighing the benefits of a hands-free experience. For those younger adults that used voice search frequently, the reason for use was convenience, specifically when multitasking. Thus, these participants reflected on using voice search in moments such as taking transportation or conducting housework, where they felt the availability of voice search outweighed the occasionally jumbled or misunderstood queries. Therefore, the benefits and participants preferences for traditional search compared to voice search were related to the usability issues they encountered when carrying out search tasks using voice such as issues with speech recognition, use of small screens, and lessened ability to compare results.

Reason for searching health information online

Online health search for most participants, both young and old, provided what participants believed was a good amount and variety of information to support their health decisions compared to other information sources (e.g., books, magazines, etc.). Participants also believed the variety of online materials represented far more people and situations compared to other information sources which they felt was important. Both older and younger adults had similar sentiments about using the Internet to support their information needs regarding their health. Both parties, older and younger, agreed that they would not solely trust the Internet for pertinent health information. However, there were also differences.
All (N = 10) older adult participants mentioned using online health search often. For example, all the older adults interviewed have access to a patient health record or portal that they frequently used to obtain health information provided to them by their medical professional. Participants shared that initially they had to adjust to checking these applications to make updates to their information or view changes in their own health record such as a change in medication or posting of test results. As this became a part of their daily routine, they shared that they would use the Internet to supplement their understanding of a diagnosis of illnesses or new medication. Participant 1 described, “It’s been a little over a year. But I was diagnosed with pulmonary fibrosis. So, you know, the doctor gave me information and stuff. But I was able to get information [online] to know where I am at, you know, giving me more in-depth information that I could read at my leisure. So, I like that all that information is handy [online], and you can find support groups and all that type of stuff, too”. Older adults appreciated the ability to seek out information beyond the scope of a clinical setting and felt that searching online was far simpler than what they perceived as waiting on a delayed response from doctors. However, as Participant 1 discussed, older adult participants often shared that they consulted the Internet as a secondary source of information compared to younger adult participants. Older adults shared that they were more often searching the Internet for information about chronic or acute diseases, or medication, thus they were more inclined to seek help from a medical professional to verify the information they found.

Younger adults felt that online health search provided a simpler way to find information and keep it in one place. Young adult participants’ discussions of online health search pertained to finding information for themselves when they were ill or providing information to others who have asked them for suggestions. Some younger participants also suggested that being able to search for information online could help mitigate feelings of stigma when attempting to find information on a topic they felt might be difficult to talk about with others. Participant 8 explained, “I think [about online health search for] something like a mental health issue, like anxiety it’s just something that’s kind of a stigma that’s hard to talk about … I think online is somewhere that you can look up this information without feeling stigmatized and without feeling shame.” Younger participants also brought up concerns about the expense related to seeking information in clinical settings. If it was not a “grave” emergency some younger participants felt it was more economical to search for something for free using a search engine. Participant 12 stated, “But sometimes you just don’t want to go out your way to talk to somebody else or sometimes it might be private and just a really simple question and you don’t want to have to go to the doctor about it […] Plus I mean Googling something, that’s free. To go into the doctor is a little expensive.” Finally, younger adults also felt searching online allowed more opportunities to find instances of what they might be going through. Participant 11 shared, “I feel searching online is better because the person I might talk to may not have complete knowledge of medicine and may not have faced the sickness I am going through right now. There might be other users who have gone through my state, and who might actually have those similar symptoms where they can confirm they had a particular disease. In that way, I can confirm that I might have a sickness or disease. It might be more confirming than getting information from one person”. Participants shared that they sometimes found solidarity with others in similar circumstances and at times it brought them comfort in addition to the ideas of different solutions that might address the health problems they are facing.

Overall, health search was categorized as advantageous only to the extent that it was taken simply as supplemental advice, and with awareness that the information could be false. Participants were comfortable searching for health information online if it was not “extremely” serious. Therefore, participants expressed that finding credible sources online that could be backed by a healthcare professional was a key part of their home health management strategies.

Challenges Searching for Health Information Online

The previous section highlighted a common theme among all participants regardless of age concerning credibility when searching health information online. In this section, we provide additional insight on the challenges participants encountered searching for health information online. All participants discussed that they were keenly aware that not all information online represented fact and they took this into consideration when filtering through online websites. Participants were also aware that information found about symptoms could not be confirmed without a formal diagnosis from a medical professional. Participant 8 echoed this sentiment, stating, “But the problem is that you don’t know if it’s accurate. For example, I can look up symptoms, but I don’t know if it’s really the same as my health issue. So, it’s really vague. I would say that you don’t know which source is trustworthy compared to other information so maybe you can know for sure if it’s accurate”. Most participants felt that filtering information for credibility was challenging due to different users’ capacity to find and process information relevant to their situation. However, older adults mentioned this concern more often. Four older adult participants acknowledged that they found it difficult to discern whether information found in an online health search was relevant or pertained to their situation. Several older adults provided as an example their experience of looking online to search for a symptom they were having only to realize after talking with a medical professional that they were led to diseases that were far worse than what they were
facing. Therefore, while the amount of information on the Internet was sometimes useful, the abundance of information on the Internet also could be problematic.

Establishing credibility when looking for online health information was varied among all participants, with some explaining a more elaborate plan for determining relevance compared to others. Younger adult Participant 11 explained how he filtered through a certain number of queries as follows, “If I have three symptoms which are related to a disease, ... I type those three symptoms and I get multiple results. I verify with each website, at least four to five websites so they will be mentioning other symptoms that could associate with that disease... So that is how I process the information from the results that I get and also maybe two or three websites that say the same thing, I will go with that result as well.” One older adult participant, Participant 6, shared that this type of filtering information can be frustrating, “It’s just general, you know, it’s just like if I would put in rashes or something, so many things would come up and I don’t want to search through this long list of things looking for that particular type of rash or whatever it is.” Older adult participants overall however were lenient to taking more structured approaches to improve the query by selecting more concise word choices or including additional words even though they acknowledged it was not always clear that it might lead to better information. For example, participant 2 an older adult participant stated, “and you kind of have to fly by the seat of the pants because you can’t tell what’s perfectly real and what’s not on the Internet. But if you get 2 or 3 different opinions or answers and theirs similar, I think you can take that as its real. I don’t trust everything on there that I read like some people do.” Like P11 and P2, other participants expressed that they often compared information on different websites to gauge credibility and were more likely to gain confidence with data if it was repeated on different websites. In addition to repetitiveness, participants also found assurances in information and assistance from health professionals.

Participants also mentioned they sometimes did not understand the information discussed in online articles due to jargon which made searching for health information online difficult. Participants mentioned issues with health literacy and terminology as one of the main reasons they were often inclined to talk to a health professional for information rather than searching online. Participants also mentioned challenges due to the overwhelming amount of online content. Participant 7 explained, “That’s a hard part. Like I said, you know you put in one subject in the computer, and it pulls up 50 things for you to look at, you know? That could be very confusing.” To accommodate challenges some participants attempted to develop strategies to narrow down the amount of information and reduce the time it takes for them to parse through. Some younger adult participants shared strategies of how they would skim through different websites rather than read them completely, because they knew not all information pertained to them. These participants felt that getting the gist of information was more important than complete details. For example, one younger participant shared, “If I go onto a website and there’s just a lot of information like an overwhelming amount, usually I’ll skim and try to I guess highlight what I’m interested in. That’s usually what I’ll do for most websites because I don’t want to read every single paragraph knowing that it doesn’t relate to what I want to know about.” Participants suggested that sometimes skimming through websites provided enough information to satisfy their reason for search. However, the amount of information returned, and the time needed to parse through it was still discussed as a source of contempt among many of the participants.

Perceptions of Using Voice for Health-Related Search

Participants varied slightly when asked their comfort level when using voice search for online health. Each group addressed specific concerns that were related to their specific perceptions of using voice technologies for health search. The dominant theme that emerged among older adult participants was that they appreciated having a wide set of health information available online regardless of the search tool. However, they felt that voice search could make it easier to access information, improve accessibility and reduce the challenges associated with generating and typing queries. Participant 6, an older adult, described the efficiency of voice-technologies as a benefit, “Just convenience is the main thing. Um, not that it takes a long time to type it in, but you have to worry about hitting the wrong buttons and all that. So, it’s so much easier just to say, ‘Hey, Google’.” Another older adult participant also reiterated the convenience of not making mistakes while typing, sharing that they felt using voice created less confusion and more of the work was being placed on devices opposed to the user. Four of the ten older adult participants had never used voice search at all for online health search but mentioned that they were not opposed to it. Participant 2 stated, “I wouldn’t be opposed to it ... I don’t have it on my phone. I suppose if I had a phone that would access it [voice search] I would be more inclined to do it”. The other three participants provided similar reasoning stating that they simply did not understand how to use voice search; however, if they were guided, they would be more likely to use it for online health search.

Younger adult participants were more critical of using voice search for online health than the older adult participants. Only two of the seven participants had used voice search in the past for online health and both participants discussed
that the negatives outweigh any benefits. Participant 12 described, “It wasn’t bad, but the experience could’ve been a lot more beneficial to me. It gave me websites that were related to what I was trying to search but they weren’t that detailed or in depth as I wanted them to be.” Being able to access sufficient information was a concern many participants had with voice search even when not conducting online health information. Participants also shared that they felt that the number of responses returned by voice search queries particularly on mobile and standalone devices are fewer, so users are not able to listen and compare multiple sources of information. Unlike older adult participants, all younger adult participants preferred to search for online health information through traditional search methods to gather a wide variety of information and to be able to manually compare results. Younger participants, unlike older adults also felt it more efficient to type and customize queries manually rather than relying on speech recognition. In addition to criticisms about voice supporting effective online health search, participants were additionally concerned about privacy and mentioned that they would be hesitant to ask about certain health topics in public and they might not be for non-health related topics. One participant therefore mentioned that they would limit online health search to situations where they were alone out of fear of revealing personal information.

Discussion

Our findings suggest that older and younger adults have different perceptions of the usefulness of voice search for consumer health information tasks. Most younger adult participants preferred traditional text search for non-health related and health related search tasks citing that voice search can make search in general more complicated and time consuming. Older adults on the other hand saw potential in the use of voice search for improving the efficiency of the search process by reducing challenges they faced with query generation and typing. Therefore, voice search for online health remains a point of contention for both older and younger adult groups. In the following suggestion we discuss potential ways forward for better utilizing voice search to support online health information search processes.

Improving Voice Interactions for Health Search

Our findings align with prior work that has explored the benefits and barriers of voice search more broadly. Prior work notes that older adults often prefer voice for its efficiency\(^1,4\) and there have been concerns about the practicality of voice search for health-related tasks\(^1\). Our findings imply that despite the growing interest and availability of consumer-facing voice technologies, open challenges exist in using these systems for health-related search tasks. From a younger adult perspective, the potential conveniences provided by voice search were counteracted by the limited set of results often returned. While voice search options are available on web browsers and return comparable results similar to traditional, text-based search, participants often equated voice search with something to be used on the go. Thus, most often participants’ discussions of online voice search were concentrated on mobile or standalone devices which have distinctive design principles that may not be ideal for some health search tasks. Thus, improving voice search for health and more generally may require a closer examination of existing conversational design principles used to guide voice assistant interactions and how these interactions can be reimagined for supporting voice health search. For example, younger adults felt that limited results returned from voice queries did not suit their needs of being able to filter through options. To accommodate a broader range of results, voice search could implement similar abilities to traditional search such as providing a broader range of options but personalize and rank the results based on prior knowledge of the user or using an approach that reveals different options based on predetermined categories. The conversational design could also be altered to provide follow-up dialog to alleviate the potential usability issues that arise when presenting long lists of options on smartphone or voice-only interfaces. Providing interactions that allow users to compare results could move forward voice search for health.

In parallel, older adults raised concerns about not knowing how to use voice search. Older adult participants’ main deterrent to voice health search was a lack of knowledge on how to use technology. Supporting discoverability or a user’s ability to successfully uncover and explore the capabilities of voice-based systems is an open challenge particularly in voice-only systems\(^37,43\). To advance voice search for health, use cases should consider options for guiding users through the voice search process. Some participants mentioned that in-person instruction might be the most suitable method for learning as opposed to the typical online and offline user guides. Guides such as these have been found problematic among older adults for promoting transparency in the past\(^16\). Thus, moving forward, it would be useful to further explore both technical and non-technical solutions to increasing older adults’ agency in voice search for health.

Addressing Trust and Safety Concerns

Trust and safety concerns mentioned by participants when using voice for health were linked to the lack of credible information available in search results and lack of knowledge of where health-related search queries were being sent.
Apart from voice health search, participants raised gauging credibility of online information for health purposes as difficult. Although participants shared different filtering methods, it was not clear whether different methods garner more verifiable information. The ability of users to identify credible information online in health-related searches is a known issue and we know that the ability to gauge credibility can be impacted by demographics and information source\(^{39,40}\). Participants’ criticisms regarding credibility implied that voice technologies provided less transparency and therefore made it more difficult to make decisions which they felt was key to facilitating online health information search.

Gauging credibility of websites prior to displaying results and then providing users with access to that information within the search result list may partially address this issue. For example, systems could rank results based on a credibility score. For voice only systems, the device could capitalize on existing approaches that announce where the information came from (e.g., “according to WebMD diabetes is …”) to provide users with data to help them decide if they want to use the information or not. However, it might be helpful to also provide information to help people gauge credible sources. For example, include information about the website itself to help users decide if they want to take information from that website (e.g., “WebMD is a credible consumer source of information based on …”). Therefore, providing some indication of the credibility of the information might help users better trust voice-based search and gauge potential risks.

**Limitations**

While we collected data until we began to see recurring themes in our data (i.e., reached saturation), our study may be limited by our participant demographics. Our participants younger and older were primarily female and most had some sort of college experience. While younger and older adults had different technology experiences, most did use technology. While some participants were familiar with voice search and had used it in the past, there were several participants that shared that they no longer used voice search or limited their use. Also, most participants that used voice for search used voice assistants which are designed to provide different types of search interactions compared to text-based search in a web browser. Therefore, our findings might be limited due to context and the specific demographic of participants. As our study is qualitative with a smaller number of participants, we do not aim for generalizability. We expect that our findings will transfer to similar groups and contexts\(^{41,42}\) however additional studies may be needed to determine relevance to other populations of users.

**Conclusions**

Our study compared experiences with text-based and voice search among older and younger adults for health and non-health related tasks. We found some differences in older and younger adults’ perceptions of the usefulness of voice for conducting health search tasks. Our findings suggest that the potential to use voice for health information search tasks still requires several improvements. A combination of increasing transparency of search results when interacting with voice-based technologies and improving voice search recognition and filtering will be essential to increase acceptance and use. Continued studies examining participant concerns, especially studying individual use, is one of our next steps to better understand the usefulness and feasibility of voice for supporting health information tasks. In addition, we will further divulge the necessary requirements for improving conversational design to accomplish transparent and efficient voice-based health search.

**Acknowledgements**

We would like to thank our participants who shared their experiences with us. We would also like to thank the Indiana University EMPOWER program for providing funds to support this research. We would also like to thank the Indiana LSAMP program (IN LSAMP) for supporting several of the undergraduate research scholars involved in this research.

**References**


35. Older Adults’ Perceptions of Intelligent Voice Assistant Privacy, Transparency, and Online Privacy Guidelines. 2020; Available from: https://www.usenix.org/conference/soups2020/presentation/bonilla
Controversial Trials First:
Identifying Disagreement Between Clinical Guidelines and New Evidence

Florian Borchert, MSc, Laura Meister, Thomas Langer, Dipl. Social Scientist, Markus Follmann, MD, MPH, MSc, Bert Arnrich, PhD, Matthieu-P. Schapranow, PhD
1 Digital Health Center, Hasso Plattner Institute, University of Potsdam, Germany
2 German Guideline Program in Oncology, German Cancer Society, Berlin, Germany

Abstract
Clinical guidelines integrate latest evidence to support clinical decision-making. As new research findings are published at an increasing rate, it would be helpful to detect when such results disagree with current guideline recommendations. In this work, we describe a software system for the automatic identification of disagreement between clinical guidelines and published research. A critical feature of the system is the extraction and cross-lingual normalization of information through natural language processing. The initial version focuses on the detection of cancer treatments in clinical trial reports that are not addressed in oncology guidelines. We evaluate the relevance of trials retrieved by our system retrospectively by comparison with historic guideline updates and also prospectively through manual evaluation by guideline experts. The system improves precision over state-of-the-art literature research strategies while maintaining near-total recall. Detailed error analysis highlights challenges for fine-grained clinical information extraction, in particular when extracting population definitions for tumor-agnostic therapies.

Introduction
With the amount of published medical evidence steadily growing, it becomes increasingly challenging for practitioners to keep up with the latest developments in their field. Clinical practice guidelines (CPGs) are designed to summarize the currently available evidence regarding specific clinical questions and provide recommendations based on well-defined and methodologically sound criteria. Despite advances in the development of living guidelines, even the most recent CPGs are ultimately static documents in a constantly evolving landscape of new evidence in the form of primary research articles, e.g., reports of randomized controlled trials (RCTs), or even unpublished results.

Finding and ranking relevant medical evidence is a well-investigated information retrieval problem. Medical literature search engines allow users to apply fine-grained search filters, based on string patterns as well as manually or automatically derived metadata, such as Medical Subject Headings (MeSH). In addition, a ranking of most relevant articles given a user query is desirable, which can be based on additional intrinsic properties of the document. In the case of a clinical trial report, this could be the study design or sample size.

While a growing amount of hand-curated and automatically derived semantic information is available for medical information retrieval, current software systems do not take into account the relationship of new evidence to the currently established CPGs, in particular as their applicability depends on the location of the user. Prospectively, new evidence may disagree with the statements in these CPGs, e.g., when an RCT presents new results that are not (yet) accounted for.

In the remainder of this work, we will refer to such evidence as controversial, which shall broadly incorporate any kind of disagreement with current CPGs. This information is relevant to a variety of audiences, for instance:

- Maintainers of CPGs, who wish to identify update signals with the potential to necessitate a CPG update
- Readers of CPGs, who need to verify whether a CPG still reflects the latest evidence
- Specialists interested in new treatments beyond CPG recommendations, e.g., for specific subgroups of patients, as it is common in precision medicine

In this work, we propose an automatic approach to identify such controversial evidence. The system is based on metadata automatically derived via natural language processing (NLP) from scientific articles and CPGs from clinical guideline repositories.
Figure 1: Overview of our approach to detect RCTs with drug interventions not covered by CPGs. Here, we consider the German CPGs in oncology in GGPONC. CPGs and RCTs from Trialstreamer are matched based on the CPG topic and the extracted population concepts of the RCT. The set of intervention concepts is compared to the drugs / chemicals mentioned in GGPONC to flag RCTs with interventions currently unmentioned in the CPG as controversial. Term expansion is performed using MeSH and SNOMED CT to account for synonymy and hyponymy.

Due to the country-specific nature of CPGs, they are typically published in their respective national language. Therefore, the underlying NLP problems are inherently multilingual. We address this issue by automatically mapping the extracted information to the Unified Medical Language System (UMLS), thereby establishing a novel link between primary and synthesized evidence across languages.

Different types of disagreement can be defined over such information using transparent rules, which enables users to reason about the validity of the identified disagreement. We demonstrate the feasibility of the proposed approach by implementing and evaluating the detection of a highly relevant form of disagreement: RCTs mentioning drug interventions not included in current CPGs, as outlined in Figure 1. To this end, we leverage two recently published resources, the Trialstreamer database of RCTs and the German Guideline Program in Oncology NLP Corpus (GGPONC).

A prototype of the system is publicly available: https://we.analyzegenomes.com/nge.

The remainder of this work is structured as follows: in the following section, we set our approach in the context of related work, followed by a description of the used datasets and components of our system. We proceed by presenting the output of the system for specific clinical indications and the evaluation in an information retrieval setting. Results are interpreted and limitations discussed thereafter, followed by a conclusion and outlook.

Related Work

Information retrieval of the medical research literature has been studied extensively in the last decades. A multitude of approaches have been proposed to filter and rank by relevancy collections of RCT reports and other medical publications.

PubMed is probably the most widely used medical search engine and allows fine-grained filtering based on text matches, MeSH terms and other metadata. Recently, PubMed has introduced a new Best Match sorting option, taking into account different metrics and user queries. The Trip database allows more advanced search options based on the Population-Intervention-Control-Outcome (PICO) framework and a ranking of articles by relevancy, such as latest and greatest. There is a wide selection of systems that enrich PubMed with additional semantic information and employ them for document retrieval, for instance using automatically extracted biomolecular entities and relations.
In practice, literature search strategies for initial CPG creation and updates are based on elaborate Boolean search queries in different literature databases, carefully hand-tailored for the specific clinical questions covered by the CPGs[1]. These queries can be continuously applied to identify update signals, i.e., research results that would necessitate an immediate update or amendment to a CPG[1]. Reviewing of the results and subsequent data extraction is done manually.

RobotReviewer is a system that automatically extracts structured information from RCT reports and predicts the risk of bias (RoB) of an RCT according to the Cochrane RoB tool[15]. Recently, Marshall et al. (2020) released Trialstreamer, combining the predictions of RobotReviewer and other components leveraging machine learning (ML) into a living database of RCTs along with their extracted metadata[6]. Based on this metadata, the Trialstreamer website allows to filter RCTs by their PICO elements and prioritize large and high-quality RCTs.

To the best of our knowledge, there is currently no system that combines information from published research results and CPGs to expose additional relevancy criteria for information retrieval.

Datasets

**Trialstreamer** is a publicly available, regularly updated database of RCTs derived from automatically screened PubMed articles and the WHO International Clinical Trials Registry Platform[9]. For this work, we focus on PubMed articles, which are included in Trialstreamer according to an automatic classification of articles that describe RCTs in humans. From the corresponding abstracts, structured metadata is extracted using different ML- and rule-based NLP methods.

For our system, we use the extracted sample size, PICO spans and PICO concepts. For the evaluation, we combine the base version of Trialstreamer and all PubMed updates up to December 28th, 2020, resulting in a dataset of around 699k RCTs. In the live version of the system, updates to Trialstreamer are automatically integrated on a weekly basis.

**GGPONC** is a metadata-enhanced text corpus based on German CPGs in oncology, currently consisting of 25 CPGs with around 1.3M tokens and more than 4k recommendations covering a diverse set of indications[7]. It is currently one of the largest publicly available text corpora of CPGs in general and of German medical text in particular. GGPONC has been automatically annotated with entity classes from different UMLS semantic types, for instance, Disorders, Procedures, Chemicals & Drugs, and Anatomy from the German subset of the UMLS. Gold-standard annotations from human experts are available for around half of the corpus.

In addition, GGPONC provides a variety of metadata for individual recommendations, e.g., timestamps, which allow us to simulate past guideline versions by excluding elements introduced after a certain point in time. GGPONC is freely available upon request[†]. New CPG versions are automatically integrated into the system upon release.

**Named Entity Recognition and Normalization**

To find drug mentions in GGPONC, we use the same dictionary-based JCoRe (i.e., UIMA-based) pipeline as in Borchert et al. (2020)[14]. In contrast to our earlier work, we configured the pipeline with a larger dictionary of substances derived from the UMLS (version 2020AB) using the JuFIT tool (v1.1)[17]. The goal of using a larger dictionary is to improve recall (sensitivity), at the expected cost of precision. Therefore, we consider all preferred English terms of the UMLS semantic type Chemicals & Drugs in addition to the German terms, yielding a dictionary of around 1.26M entries compared to only 34.550 from the German UMLS subset. With this extended dictionary, recall for the recognition of Chemicals & Drugs measured on the human-annotated subset of GGPONC increases from 0.600 to 0.788, whereas precision decreases from 0.917 to 0.520. Due to the nature of this dictionary-based approach, all extracted entities are already linked to UMLS CUIs (concept unique identifiers).

The Trialstreamer database already contains automatically derived metadata relating to PICO elements. Entities within PICO spans are linked to UMLS CUIs using a re-implementation of Metamap Lite[18]. Recall for PICO concept extraction under relaxed comparison reported by Marshall et al. (2020) is relatively high (0.85 for intervention and 0.78 for population), while precision is rather low (0.57 for intervention and 0.30 for population)[5].

[†]https://www.leitlinienprogramm-onkologie.de/projekte/ggponc-english/
Identifying Controversial RCTs in Trialstreamer

The current version of our system enables the detection of controversial RCTs in Trialstreamer by linking the extracted PICO concepts to CPGs from GGPONC as outlined in Figure 1. In the following, we describe the rule-based algorithm that detects RCTs concerning drugs unmentioned in current CPGs. Other types of rules can be implemented similarly.

For all RCTs in Trialstreamer, we identify those with intervention CUIs contained in the large UMLS-based dictionary of drugs and chemicals described earlier and retain only such trials where at least one intervention is a drug or chemical, as opposed to other kinds of treatments. For each guideline topic, e.g., lung cancer, we perform a term expansion step to identify synonyms and hyponyms, such as pulmonary neoplasm, lung adenocarcinomas, or bronchogenic carcinoma, which we refer to as topic concepts. This expansion was performed based on the MeSH and SNOMED CT terminologies using PyMedTermino and the entity linking module from scispaCy.

To obtain all potentially relevant RCTs for an existing CPG topic, we filter Trialstreamer by the automatically extracted population CUIs that overlap with the set of topic concepts. For each of these topic-related RCTs, we determine whether at least one of the intervention CUIs is not contained in the complete set of drug CUIs extracted from the corresponding CPG at a particular time point. Such RCTs are flagged as controversial. Term expansion is also performed on the level of interventions, to account for mentions of drug classes, i.e., an RCT mentioning a more general term (e.g., protein kinase inhibitor) should not be considered controversial if a narrower term (e.g., afatinib) is already included in the CPG. This step also accounts for synonymy due to the use of experimental, non-proprietary, and trade names of drugs.

This simple rule over the extracted metadata is completely transparent to the user. Moreover, it is easily extensible and adaptable to different requirements, e.g., when a different trade-off of precision and recall is desired, and can be easily implemented within medical search engines. We provide a prototypical user interface, displayed in Figure 2, that allows users to browse RCTs per CPG topic and to display only controversial results, in addition to filtering based on the publication year, sample size, and free-text matches. It should be noted that even simply filtering by sample size is not available in typical literature search engines and only enabled by the NLP-derived information in Trialstreamer.
Figure 3: Development of the number of RCTs for the top 15 drug names not mentioned in the previous CPG for two topics and update intervals. We show the initial number of RCTs in the year of the previous CPG update as well as the cumulated number of RCTs until the current update. Drugs marked as incl. have eventually been included in this last update, whereas drugs marked as not incl. are (still) not mentioned in the current CPG version.

Evaluation

Retrospective Scenario Using the temporal metadata in GGPONC, we simulate CPG versions at a past point in time by manually identifying and removing all drug mentions from the CPGs that were first included after that point in time. Two CPGs and simulated versions are investigated: a 2010 version of the lung cancer (LC) CPG, which has received an update in 2018, as well as a 2017 version of the renal cell carcinoma (RCC) CPG, updated in 2020. For these two topics, we process all RCTs from Trialstreamer and flag them as controversial if at least one of their interventions is a drug not yet included in the historic CPG.

In Figure 3, we show the drugs with highest numbers RCTs identified in this way. At the points in time the CPGs were issued, there are only very few interventions mentioned in RCTs that are going to be included in future CPG updates — the respective CPG can be considered up-to-date. In the following time preceding the next update, evidence relating to the most commonly mentioned drugs has accumulated. However, for both CPGs, there is a noticeable tail of drugs investigated in clinical trials that do not end up in the respective CPG.

Retrospective Evaluation of Retrieval Performance In the considered scenario, where our goal is to find interventions not currently part of CPGs but with the potential to become CPG-relevant in the future, a typical requirement is near-perfect recall, generally at the cost of precision. Using the controversial flag as an additional filtering criterion for retrieving RCTs from PubMed, we evaluate retrieval performance in terms of (1) precision with respect to retrieved documents actually describing an RCT with a newly CPG-relevant drug, as well as (2) recall with respect to the proportion of retrieved drugs from the set newly CPG-relevant drugs. The results are given in Table 1. Note that retrospective assessment of recall in terms of documents is actually not possible based on our data, since a particular CPG update cannot be reliably attributed to a single RCT. This would require manual reconstruction of the literature screening process, which might be partially based on existing systematic reviews not included in Trialstreamer.
Table 1: Retrospective evaluation of retrieval performance under different filters evaluated against past CPG updates. The baseline strategies are filters by Year (Y), by drug interventions only (I\textsubscript{Drug}), by matching the study population to any sub type of Cancer (P\textsubscript{Cancer}), or to the respective CPG topic (P\textsubscript{Topic}). We compare these to advanced filters based on the sample size of n from Trialstreamer, and the controversial tag from our method (Controv.). We identify all RCTs in Trialstreamer matched by these criteria over the period between two consecutive updates and report precision (Pr.) in terms of documents and recall (Rec.) in terms of retrieved drugs newly mentioned in the current CPG version. Very high precision can be obtained by our method based on the controversial tag, in particular in combination with a filter based on the sample size (n > 100), while maintaining the same recall as the most specific baseline strategy.

Prospective Evaluation In a real-world deployment, we want our system to detect controversial trials based on the current CPG versions, in contrast to historic versions used in the retrospective evaluation. Therefore, to evaluate the system prospectively, we consider the current versions of the German CPGs on Ovarian Cancer (OC) and Malignant Melanoma (MM), both with a last update in the early 2020, and identify controversial trials with respect to these topics in 2020. As the interventions relevant for the next CPG version are still unknown in the scenario, the output of the system was manually evaluated by guideline experts from the German Cancer Society.

The manual evaluation results are shown in Table 2. 41 RCTs are found by our system for Ovarian Cancer and 21 for Malignant Melanoma in, resulting in a total of 62 articles from 2020 used in this evaluation. 23 of these have been categorized to have direct potential to be relevant for future CPG versions, corresponding to a precision of .355, when applying the same standard as in the retrospective evaluation. When we also consider early-stage RCTs as relevant and also ones reporting negative results, precision increases to .645. In effect, only 35.5% of results are actually clearly irrelevant, because of errors in downstream components of the system. This number is very low compared to usual Boolean searches in PubMed. We will analyze the different error conditions in the next section.
### Table 2: Prospective evaluation for the system configurations based on the current German CPGs on Ovarian Cancer and Malignant Melanoma, both from early 2020. For all identified interventions, we report the expert decision regarding the potential relevance for a future CPG update. We also denote if an RCT does not meet the criteria for CPG inclusion due to an early phase or negative results. If there is more than one publication per intervention, this number is denoted in brackets. Actual false positives are designated in the last column, with the error type distinguished by origin between ontology incompleteness (O) and errors in Trialstreamer: RCT document classification errors (TS-D), intervention extraction errors (TS-I), and population extraction errors (TS-P).

**Notes:**
- †Description of the trial design only.
- *Relevant for the population, but not the underlying clinical questions used during CPG development.
- **No or unclear relevance in the EU / Germany.

#### Error Analysis

**Phase I/II Trials, Negative Results, and Study Withdrawal**  Phase I or phase II trials, e.g., for TRC105 in RCC or cabazitaxel in OC, are not immediately relevant for CPG inclusion. Negative results in such phase I or II trials or study withdrawal occur frequently. These types of results have to be considered as false positives with respect to the goal of identifying CPG update signals, and are counted as such in the retrospective evaluation. However, they can be of great interest to other users of the system. Being able to distinguish the particular type and design of an RCT automatically would be a useful feature in a future version of the system, but would require a more elaborate semantic understanding of study contents through NLP. Negative results for drugs included in current CPGs would constitute a different type of disagreement that we would like to be able to detect.

**Misalignment of Topics and Populations**  When the scope of a CPG is not fully defined by the population alone, some results might be relevant for the topic but not (yet) covered as a clinical question in the CPG. For instance, supportive therapy with royal jelly or the use of local anesthetics might be relevant for the considered population, but covered in a separate CPG on such cross-cutting concerns. Addressing this problem comprehensively would require a complete formal specification of the clinical questions underlying each CPG.
**False Negatives** For this type of error, drugs would be relevant, e.g., for future CPG versions, but we retrieve no RCTs mentioning them (lower recall). In the retrospective evaluation, recall is < 1 only for LC after applying a population filter, because the drug dabrafenib, newly mentioned in the 2018 guideline, was missed. Dabrafenib could not be found in Trialstreamer after filtering by population, as all RCTs on the drug mention more general population terms, such as solid tumors, which is not captured by our term expansion strategy.

**False Positives** For this type of error, RCTs are retrieved that mention drug names which are not included in the future CPG version (lower precision). Sources of false positives, by example of the drugs depicted in Figure 3 and Table 2 are:

- Ontology incompleteness (O): if no mapping of a concept exists in the UMLS for the language of the CPG (here German), it would not be extracted from the current CPG and therefore not be considered as already mentioned. Incomplete terminologies in the UMLS can also be a source of false negatives, e.g., when novel drugs have not yet been curated. Such a false negative, however, did at least not occur in the retrospective evaluation.
- Document classification errors (TS-D): errors introduced due to a misclassification of a publication as an RCT. In one case, a narrative review on denosumab has been misclassified as an RCT.
- Intervention extraction errors (TS-I): for example, rapamycin is extracted as an intervention in Trialstreamer and normalized to sirolimus, although it only occurs as part of a larger entity mammalian Target of rapamycin. Given the high reported accuracy of Trialstreamer with respect to the extraction of interventions, this error type is only encountered occasionally. This class of errors includes incorrect abbreviation expansion, as in the case of DCA, which was considered as a chemical but supposed to mean decision curve analysis in the RCT.
- Population extraction errors (TS-P): in some cases, the study population has been extracted incorrectly in Trialstreamer. These problems can be rather obvious, e.g., for an RCT with breast cancer patients performed at the Center for Familial Breast and Ovarian Cancer. Other instances are more subtle, e.g., when the population malignant melanoma is assigned to an RCT with choroidal melanoma patients in Trialstreamer.

**Discussion**

**Limitations** Our detailed error analysis highlights limitations of the system as currently implemented. The initial population matching strategy is currently unable to account for drugs like dabrafenib, whose target population is based on molecular characteristics rather than specific tumor entities, i.e., tumor-agnostic therapies.

In general, the performance of our system strongly relies on various downstream components, in particular the ML-based components used to populate Trialstreamer as well as the dictionary-based information extraction from German CPGs. Reliance on currency and completeness of the UMLS is problematic in particular for new concepts and low-resource language communities. While the NER step could be solved using an ML-based approach, matching of entities across languages without access to a multilingual ontology will be challenging. It further needs to be investigated if truly controversial treatments will be detected by current NER solutions, as they might not be adequately represented in terminologies or training data.

An assessment of recall and an error analysis of false negatives in the prospective setting is still missing, as it would require a screening of all articles to account for false negatives in the classification of RCTs by Trialstreamer as well as incompleteness of the UMLS with respect to novel drugs. In a future evaluation, we will apply the system in parallel to an ongoing major update of the German LC CPG, including a full literature screening for relevant RCTs.

The current binary scheme, which flags each RCT as either controversial or not, could be enhanced by softer relevancy criteria. These should account for the inherently probabilistic outputs of downstream NLP components and incorporate relationships to multiple, potentially disagreeing CPGs. Using such criteria would enable the implementation of a true ranking of results and the use of ranking-based evaluation criteria, such as P@K or AUC.

**Other Types of Evidence** While RCTs are the gold standard in evidence-based medicine, they are by far not the only source of medical evidence. For many uses cases, interventions investigated in ongoing RCTs or with unreported results (publication bias) will be relevant and therefore clinical trial registers, such as ClinicalTrials.gov should be
considered. This information will be easier to incorporate in the system, as data in these registers is usually available at least in a semi-structured format, and already included in Trialstreamer. Other types of published research, such as case reports, are not included in Trialstreamer and will require custom NLP solutions for information extraction and normalization.

In addition, results which were not published in medical journals and other types of so-called grey literature can be of interest to some users. Incorporating these sources of evidence is likely to yield much more controversial information compared to published RCTs. However, these sources will contain a substantial proportion of completely irrelevant information. They are also expected to be more challenging to process via NLP due to large linguistic heterogeneity.

**Other Controversial Results** There are many types of controversial evidence not yet covered by our methodology. A different type of intervention would be addressed by the UMLS semantic type Procedures, e.g., for radiation therapy or other nonpharmaceutical interventions. Here, we expect cross-lingual matching to be more challenging compared to drug names, which are mostly single-token proper nouns.

Other types of disagreement could occur on the level of outcomes, where new results regarding efficacy and safety of a CPG-recommended drug can influence decision-making. Extraction quality with respect to outcomes is lowest for all PICO elements extracted in Trialstreamer, so a detailed investigation and improved NLP solutions are necessary.

**Conclusion and Future Work**

In this work, we presented a system that allows to detect controversial trials, i.e., ones that disagree with current clinical practice guidelines. The system relies on the NLP-based extraction of metadata from RCTs and CPGs and simple rules to identify disagreement based on this data. While the evaluation results are encouraging in an information retrieval setting, there is ample opportunity for improvement.

With more and more drugs targeting specific genetic alterations, this information should be incorporated into the population matching strategy, in particular for other types of users, e.g., participants of molecular tumor boards. Information extraction of molecular entities has been studied extensively by the BioNLP community, albeit rarely in the domain of RCTs. Such an extension would make the system also useful for curators of clinical evidence in precision oncology knowledge bases.

To enable the implementation of rules for other types of disagreements, future work will focus on the inclusion of information regarding outcomes within the PICO framework as well as different published and non-published types of evidence. To further improve system performance in the investigated multi-lingual setting, we will need to improve clinical NLP methods for languages other than English, which are still restricted by the shortage of publicly available research dataset and, in comparison to the English language community, a lack of modern ML-based information extraction solutions.

A prototype of our system is online available for evaluation at [https://we.analyzegenomes.com/ng](https://we.analyzegenomes.com/ng). We will continuously improve the prototype to allow users to find other types of disagreement, incorporate different kinds of evidence and extend the scope to other CPGs from other medical fields and countries.

**Acknowledgements**

This work was partially supported by a grant of the German Federal Ministry of Research and Education (01ZZ1802).

**References**

Bias Assessment and Correction in Machine Learning Algorithms: A Use-Case in a Natural Language Processing Algorithm to Identify Hospitalized Patients with Unhealthy Alcohol Use

Marissa Borgese, MS1, Cara Joyce, PhD2, Emily E. Anderson, MPH, PhD2, Matthew M. Churpek, MD, MPH, PhD3, Majid Afshar, MD, MSCR2,3

1Loyola University Chicago Stritch School of Medicine, Maywood, IL; 2Loyola University Chicago, Chicago, IL; 3University of Wisconsin, Madison, WI

Abstract

Unhealthy alcohol use represents a major economic burden and cause of morbidity and mortality in the United States. Implementation of interventions for unhealthy alcohol use depends on the availability and accuracy of screening tools. Our group previously applied methods in natural language processing and machine learning to build a classifier for unhealthy alcohol use. In this study, we sought to evaluate and address bias through the use-case of our classifier. We demonstrated the presence of biased unhealthy alcohol use risk underestimation among Hispanic compared to Non-Hispanic White inpatients, 18- to 44-year-old compared to 45 years and older medical/surgical inpatients, and Non-Hispanic Black compared to Non-Hispanic White medical/surgical inpatients. We further showed that intercept, slope, and concurrent intercept and slope recalibration resulted in minimal or no improvements in bias-indicating metrics within these subgroups. Our results exemplify the importance of integrating bias assessment early into the classifier development pipeline.

Introduction

An estimated 14 million United States adults met criteria for an alcohol use disorder in 2018, with nearly 1 in 10 adult alcohol users affected1. Unhealthy alcohol use represents a major economic burden and causal factor in cases of traumatic injury, liver disease, and other noncommunicable diseases resulting in death2,3. Further, alcohol-related disorders consistently rank in the top ten most common non-maternal diagnoses responsible for inpatient stays among adults under 44 years old4 and individuals with alcohol-related disorders are more likely to return to the hospital within two weeks of being discharged compared to unaffected individuals5. Screening tools for unhealthy alcohol use can effectively identify individuals who will benefit from interventions that decrease alcohol use-related morbidity and mortality6,7. Accordingly, the successful implementation of treatment depends on the availability and accuracy of such screening tools.

Unhealthy alcohol use screening in the hospital setting currently involves the use of a single screening question or a standardized questionnaire such as the Alcohol Use Disorders Identification Test (AUDIT)8. Adding self-report questionnaires into the clinical workflow requires additional time and resources that may limit their utility. Screening methods that utilize notes from the electronic health record (EHR) captured during routine care are promising alternative approaches to identify likely cases of unhealthy alcohol use9.

For clinical decision support tools, there is growing concern about the potential for bias in machine learning (ML)-based automated approaches. There have been several notable publications discussing unintended bias in ML tools across several fields, with consequences ranging from underprediction of health risk in Black patients by a widely used commercial insurance algorithm10 to increased error rate in speech recognition for Black speakers11 and undervaluing of female job candidates12. In medicine, minimizing bias in ML-based clinical decision support is critical to avoiding downstream harm and the exacerbation of existing healthcare disparities.

Our group previously applied methods in natural language processing (NLP) and ML to build a classifier for unhealthy alcohol use from notes collected in the EHR that offers adequate sensitivity for screening patients in the acute care setting9. The classifier was trained on EHR notes from patients with a primary admission for trauma and has since been validated using data from non-trauma inpatient hospitalizations13. In this study, we seek to evaluate and address bias through the use-case of our previously developed unhealthy alcohol use NLP classifier. Our aims are the following: (1) to assess for bias in the NLP classifier via examination of subgroups of age, sex, and race/ethnicity; and (2) to determine if recalibration among subgroups affected by model bias can mitigate the bias in the NLP classifier’s screening performance.
Methods

Unhealthy Alcohol Use Classifier

The unhealthy alcohol use NLP classifier previously published by our group\(^9\) was developed using concept unique identifiers (CUIs) derived from linguistic processing of clinical notes, with AUDIT scores ≥5 and ≥8 points as the reference standard for unhealthy alcohol use in women and men, respectively. Clinical notes available from within 24 hours of presentation to the Emergency Department were scanned for Unified Medical Language System entity mentions, which were mapped to CUIs. Hyperparameters were tuned to maximize the area under the receiver operating characteristic curve (AUC ROC) using 10-fold cross-validation. The final classifier retained 16 CUI features and achieved an average AUC ROC of 0.78 (95% CI 0.68-0.89).

Patient Setting and Data

Our data consisted of two cohorts: one from the original development paper using trauma patient encounters in the Emergency Department (n=1,326) and one from the validation cohort of an independent group of hospitalized patients in the medical/surgical wards (n=999).

Our trauma dataset included consecutive patients seen at Loyola University Medical Center’s (LUMC) Trauma Center who received screening for unhealthy alcohol use using the AUDIT questionnaire between April 2013 and November 2016. Results from the AUDIT were used as the reference standard for labeling cases of unhealthy alcohol use, with scores ≥5 and ≥8 points indicating unhealthy alcohol use in women and men, respectively\(^8\). The trauma cohort reflects the pooled training and internal validation data used for the initial development and validation of the unhealthy alcohol use NLP classifier\(^9\).

Our inpatient dataset consisted of a convenience sample of patients who presented to LUMC for non-trauma inpatient hospitalizations between January 2007 and September 2017. Unhealthy alcohol use was identified via chart review by a trained annotator following standardized criteria previously described\(^13\) with an oversampling of at-risk patients to provide a more balanced dataset for evaluating cases. The inpatient cohort reflect the data used for external validation of the NLP classifier\(^13\).

Predicted probabilities of unhealthy alcohol use were generated for all encounters in both cohorts, and the optimal cutoffs were determined using the Youden index maximization method. The optimal predicted probability cutoffs for identifying unhealthy alcohol use were ≥0.46 and ≥0.53 for the trauma and inpatient cohorts, respectively. In the trauma cohort, the classifier had a sensitivity and specificity of 61% (95% CI 56%-67%) and 78% (95% CI 75%-80%), respectively. In the external validation inpatient cohort, the classifier had a sensitivity and specificity of 86% (95% CI 83%-89%) and 82% (95% CI 78%-85%), respectively. Descriptive statistics for patient characteristics, including age, sex, ethnicity, and race, were presented for both cohorts. All analyses were performed using R version 3.6.0 (R Core Team, 2019).

Bias Assessment

We assessed for bias in the trauma and inpatient cohorts independently and by age group, sex, and race/ethnicity. For the purposes of bias assessment and correction, age was divided into two groups \textit{a priori} based on census age groups and sample sizes: 18 to 44 years old and 45 years and older. Race/ethnicity was divided into three groups: Hispanic, Non-Hispanic Black, and Non-Hispanic White. The reference standard labels and predicted labels for unhealthy alcohol use (present/absent) were used to calculate the number of false positives (FP), true positives (TP), false negatives (FN) and true negatives (TN) within each cohort and within subgroups of each cohort. We then calculated the bias assessment metrics of interest, which included false discovery rate (FDR; FP/[FP+TP]), false positive rate (FPR; FP/[FP+TN]), false omission rate (FOR; FN/[FN+TN]), and false negative rate (FNR; FN/[FN+TP]). All bias metrics were calculated as described by Saleiro, Kuester\(^14\) and 95% exact binomial confidence intervals were calculated for all metrics.

High FDR and/or FPR values were used as indicators of bias towards overestimation of risk within subgroups; high FOR and/or FNR values were used as indicators of bias towards underestimation of risk within subgroups. Combinations of high FDR and/or FPR values with high FOR and/or FNR values were interpreted as overall reduced model accuracy within subgroups.
Bias Correction

For the purposes of developing an adequate screening tool with the NLP classifier, we focused our bias correction efforts on maintaining high sensitivity with few false negative results, so subgroups with high FOR and/or FNR values were targeted. Model calibration of intercept and slope was assessed using calibration plots in accordance with the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) guidelines. Model recalibration was implemented for each cohort subgroup that had evidence of poor calibration. Predicted probabilities within affected subgroups were used to calculate linear predictors in a logistic regression classifier, which were subsequently used for model recalibration via intercept re-estimation, slope re-estimation, and concurrent intercept and slope re-estimation. These recalibration methods were chosen to minimize the number of re-estimated parameters, control familywise error rate, and to avoid overfitting given the relatively small sample sizes available among select subgroups within both cohorts. The original NLP classifier and all three recalibrated classifiers were compared via calibration plots and scaled Brier scores with bootstrapped 95% confidence intervals (1,000 iterations). A model's scaled Brier score reflects the mean square error of probability predictions, or Brier score, scaled by its maximum score such that values range from 0% to 100% and 100% indicates optimal performance. The bias-indicating metrics were also recalculated for affected subgroups using predicted probabilities and a predicted probability threshold determined via Youden index maximization from each of the three recalibration methods.

Results

Patient Characteristics

Our trauma cohort consisted of 1,326 encounters from 1,309 patients and our inpatient cohort consisted of 999 encounters from 856 patients. Age at first encounter, sex, ethnicity, and race distributions were similar in the trauma and inpatient cohorts (Table 1).

Table 1. Patient characteristics by cohort.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Trauma cohort (n=1,309)</th>
<th>Inpatient Cohort (n=856)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age, years, median (IQR)</td>
<td>45 (28-62)</td>
<td>50 (39-59)</td>
</tr>
<tr>
<td>Sex, male, n (%)</td>
<td>921 (70)</td>
<td>567 (66)</td>
</tr>
<tr>
<td>Hispanic, n (%)</td>
<td>231 (18)</td>
<td>109 (13)</td>
</tr>
<tr>
<td>Race, n (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>American Indian</td>
<td>1 (0)</td>
<td>2 (0)</td>
</tr>
<tr>
<td>Asian</td>
<td>10 (1)</td>
<td>1 (0)</td>
</tr>
<tr>
<td>Black</td>
<td>323 (25)</td>
<td>261 (30)</td>
</tr>
<tr>
<td>White</td>
<td>778 (59)</td>
<td>502 (59)</td>
</tr>
<tr>
<td>Other</td>
<td>188 (14)</td>
<td>86 (10)</td>
</tr>
</tbody>
</table>

IQR = interquartile range.

Trauma Cohort Bias Assessment

The case-rate of unhealthy alcohol use was 23% (n=305) across all encounters in the trauma cohort (Table 2). The FPR and FOR were higher among patients 18 to 44 years old compared to patients 45 years and older, indicating reduced model accuracy among adults 44 years and younger in the trauma cohort. The FPR and FOR were also higher among male patients compared to female patients, indicating reduced model accuracy among adult patients in the trauma cohort. The FOR was higher among Hispanic patients (FOR=0.22, 95% CI 0.16-0.30) compared to Non-Hispanic White patients (FOR=0.11, 95% CI 0.08-0.14), indicating biased underestimation of unhealthy alcohol use risk among Hispanic patients compared to Non-Hispanic White patients in the trauma cohort by the NLP classifier.

Inpatient Cohort Bias Assessment

The case-rate of unhealthy alcohol use was 58% (n=579) in the inpatient cohort (Table 3). The FOR and FNR were higher among patients 18 to 44 years old (FOR=0.32, 95% CI 0.25-0.40; FNR=0.20, 95% CI 0.15-0.26) compared to patients 45 years and older (FOR=0.12, 95% CI 0.09-0.17; FNR=0.10, 95% CI 0.07-0.14), indicating biased underestimation of unhealthy alcohol use risk among adults 44 years and younger compared to adults 45 years and older in the inpatient cohort. The FOR and FNR were also higher among Non-Hispanic Black patients (FOR=0.28,
Table 2. Bias report for the trauma cohort.

<table>
<thead>
<tr>
<th>Population</th>
<th>n</th>
<th>Unhealthy Alcohol Use Prevalence</th>
<th>FDR</th>
<th>FPR</th>
<th>FOR</th>
<th>FNR</th>
</tr>
</thead>
<tbody>
<tr>
<td>All encounters</td>
<td>1,326</td>
<td>0.23</td>
<td>0.55</td>
<td>0.22</td>
<td>0.13</td>
<td>0.39</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.50-0.60)</td>
<td>(0.20-0.25)</td>
<td>(0.11-0.15)</td>
<td>(0.33-0.44)</td>
</tr>
<tr>
<td>Age group</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18 to 44 years</td>
<td>654</td>
<td>0.28</td>
<td>0.60</td>
<td>0.33</td>
<td>0.20</td>
<td>0.43</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.53-0.66)</td>
<td>(0.29-0.37)</td>
<td>(0.16-0.24)</td>
<td>(0.36-0.50)</td>
</tr>
<tr>
<td>45 years and older</td>
<td>672</td>
<td>0.18</td>
<td>0.47</td>
<td>0.13</td>
<td>0.08</td>
<td>0.33</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.39-0.55)</td>
<td>(0.11-0.16)</td>
<td>(0.06-0.10)</td>
<td>(0.24-0.42)</td>
</tr>
<tr>
<td>Sex</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>394</td>
<td>0.16</td>
<td>0.56</td>
<td>0.15</td>
<td>0.08</td>
<td>0.37</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.45-0.66)</td>
<td>(0.11-0.19)</td>
<td>(0.05-0.11)</td>
<td>(0.25-0.50)</td>
</tr>
<tr>
<td>Male</td>
<td>932</td>
<td>0.26</td>
<td>0.55</td>
<td>0.26</td>
<td>0.16</td>
<td>0.39</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.49-0.60)</td>
<td>(0.23-0.30)</td>
<td>(0.13-0.19)</td>
<td>(0.33-0.46)</td>
</tr>
<tr>
<td>Race/ethnicity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hispanic</td>
<td>232</td>
<td>0.33</td>
<td>0.46</td>
<td>0.22</td>
<td>0.22</td>
<td>0.46</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.35-0.58)</td>
<td>(0.16-0.30)</td>
<td>(0.16-0.30)</td>
<td>(0.35-0.58)</td>
</tr>
<tr>
<td>Non-Hispanic Black</td>
<td>324</td>
<td>0.21</td>
<td>0.65</td>
<td>0.29</td>
<td>0.14</td>
<td>0.42</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.55-0.73)</td>
<td>(0.23-0.35)</td>
<td>(0.09-0.19)</td>
<td>(0.30-0.55)</td>
</tr>
<tr>
<td>Non-Hispanic White</td>
<td>698</td>
<td>0.21</td>
<td>0.54</td>
<td>0.19</td>
<td>0.11</td>
<td>0.36</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.46-0.61)</td>
<td>(0.16-0.23)</td>
<td>(0.08-0.14)</td>
<td>(0.29-0.45)</td>
</tr>
</tbody>
</table>

FDR=false discovery rate; FPR=false positive rate; FOR=false omission rate; FNR=false negative rate.

95% CI 0.21-0.37; FNR=0.21, 95% CI 0.15-0.27) compared to Non-Hispanic White patients (FOR=0.13, 95% CI 0.09-0.18; FNR=0.10, 95% CI 0.07-0.14), indicating underestimation of risk among Non-Hispanic Black patients compared to Non-Hispanic White patients in the inpatient cohort.

Table 3. Bias report for the inpatient cohort.

<table>
<thead>
<tr>
<th>Population</th>
<th>n</th>
<th>Unhealthy Alcohol Use Prevalence</th>
<th>FDR</th>
<th>FPR</th>
<th>FOR</th>
<th>FNR</th>
</tr>
</thead>
<tbody>
<tr>
<td>All encounters</td>
<td>999</td>
<td>0.58</td>
<td>0.13</td>
<td>0.18</td>
<td>0.19</td>
<td>0.14</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.11-0.16)</td>
<td>(0.15-0.22)</td>
<td>(0.16-0.23)</td>
<td>(0.11-0.17)</td>
</tr>
<tr>
<td>Age group</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18 to 44 years</td>
<td>360</td>
<td>0.66</td>
<td>0.10</td>
<td>0.17</td>
<td>0.32</td>
<td>0.20</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.06-0.15)</td>
<td>(0.11-0.25)</td>
<td>(0.25-0.40)</td>
<td>(0.15-0.26)</td>
</tr>
<tr>
<td>45 years and older</td>
<td>639</td>
<td>0.54</td>
<td>0.15</td>
<td>0.19</td>
<td>0.12</td>
<td>0.10</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.12-0.20)</td>
<td>(0.15-0.24)</td>
<td>(0.09-0.17)</td>
<td>(0.07-0.14)</td>
</tr>
<tr>
<td>Sex</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>343</td>
<td>0.53</td>
<td>0.14</td>
<td>0.16</td>
<td>0.18</td>
<td>0.16</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.09-0.20)</td>
<td>(0.10-0.22)</td>
<td>(0.12-0.24)</td>
<td>(0.11-0.22)</td>
</tr>
<tr>
<td>Male</td>
<td>656</td>
<td>0.61</td>
<td>0.13</td>
<td>0.20</td>
<td>0.20</td>
<td>0.13</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.10-0.17)</td>
<td>(0.15-0.26)</td>
<td>(0.16-0.26)</td>
<td>(0.10-0.17)</td>
</tr>
<tr>
<td>Race/ethnicity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hispanic</td>
<td>117</td>
<td>0.59</td>
<td>0.18</td>
<td>0.25</td>
<td>0.27</td>
<td>0.19</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.09-0.29)</td>
<td>(0.14-0.40)</td>
<td>(0.15-0.41)</td>
<td>(0.10-0.30)</td>
</tr>
<tr>
<td>Non-Hispanic Black</td>
<td>306</td>
<td>0.60</td>
<td>0.15</td>
<td>0.21</td>
<td>0.28</td>
<td>0.21</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.10-0.21)</td>
<td>(0.14-0.29)</td>
<td>(0.21-0.37)</td>
<td>(0.15-0.27)</td>
</tr>
<tr>
<td>Non-Hispanic White</td>
<td>538</td>
<td>0.56</td>
<td>0.13</td>
<td>0.16</td>
<td>0.13</td>
<td>0.10</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.09-0.17)</td>
<td>(0.12-0.22)</td>
<td>(0.09-0.18)</td>
<td>(0.07-0.14)</td>
</tr>
</tbody>
</table>

FDR=false discovery rate; FPR=false positive rate; FOR=false omission rate; FNR=false negative rate.
**Trauma Cohort Bias Correction**

For the FOR and FNR metrics, the strongest bias in the trauma cohort with significant disparity was between the Hispanic and Non-Hispanic White subgroups. The original calibration primarily demonstrated overestimation of risk across the quintiles of predicted probabilities (Figure 1). Efforts at recalibration show improvement in movement towards the loess curve for perfect calibration. The concurrent intercept and slope re-estimation model resulted in the most significant improvement (Figure 1). The scaled Brier scores also reflected improvements in accuracy with intercept re-estimation (12%, 95% CI 9%-14%), slope re-estimation (15%, 95% CI 10%-21%), and concurrent intercept and slope re-estimation (17%, 95% CI 12%-21%) over the original model (7%, 95% CI 2%-10%).

The optimal predicted probability cutoffs were ≥0.33, ≥0.37, and ≥0.33 for the intercept re-estimation, slope re-estimation, and concurrent intercept and slope re-estimation models in the Hispanic subgroup, respectively. Despite improvements in model accuracy with recalibration, biased underestimation of unhealthy alcohol use risk (as indicated by a high FOR) persisted in the intercept re-estimation (FOR=0.21, 95% CI 0.15-0.29), slope re-estimation (FOR=0.21, 95% CI 0.15-0.28), and concurrent intercept and slope re-estimation (FOR=0.21, 95% CI 0.15-0.29) models compared to the original model (FOR=0.22, 95% CI 0.16-0.30). Minimal changes in FDR, FPR, and FNR were observed after recalibration.

![Figure 1](image.png)

**Figure 1.** Observed versus predicted probability of unhealthy alcohol use among Hispanic patients in the trauma cohort according to the (A) original versus intercept re-estimation model, (B) original versus slope re-estimation model, and (C) original versus concurrent intercept and slope re-estimation model. Five bins were used for each graph.

**Inpatient Cohort Bias Correction**

For the FOR and FNR metrics, the strongest bias in the inpatient cohort with significant disparity was between the age groups. Our recalibration methods failed to produce noticeable improvements in predictive accuracy of the NLP classifier within the 18- to 44-year-old subgroup of the inpatient cohort (Figure 2). The scaled Brier scores also indicated a lack of improvement in model accuracy with intercept re-estimation (41%, 95% CI 37%-44%), slope re-estimation (44%, 95% CI 40%-49%), and concurrent intercept and slope re-estimation (44%, 95% CI 40%-49%) over the original model (41%, 95% CI 37%-44%).

The optimal predicted probability cutoffs were ≥0.47, ≥0.43, and ≥0.43 for the intercept re-estimation, slope re-estimation, and concurrent intercept and slope re-estimation models in the 18- to 44-year-old subgroup, respectively. Using the recalibrated predicted probabilities and cutoffs, there was an equal, moderate decrease in FNR (FNR=0.12, 95% CI 0.08-0.17) across all recalibration methods when compared to the original model (FNR=0.20, 95% CI 0.15-0.26). We observed minimal changes in FDR, FPR, and FOR with recalibration.

As was the case in the 18- to 44-year-old subgroup, no noteworthy improvements in model accuracy were observed with model recalibration within the Non-Hispanic Black subgroup of the inpatient cohort (Figure 3). The scaled Brier scores similarly indicated a lack of improvement in model accuracy with intercept re-estimation (41%, 95% CI 37%-44%), slope re-estimation (43%, 95% CI 38%-48%), and concurrent intercept and slope re-estimation (43%, 95% CI 39%-48%) over the original model (40%, 95% CI 37%-44%).
Figure 2. Observed versus predicted probability of unhealthy alcohol use among 18- to 44-year-old patients in the inpatient cohort according to the (A) original versus intercept re-estimation model, (B) original versus slope re-estimation model, and (C) original versus concurrent intercept and slope re-estimation model. Five bins were used for each graph.

The optimal predicted probability cutoffs were $\geq 0.55$, $\geq 0.63$, and $\geq 0.59$ for the intercept re-estimation, slope re-estimation, and concurrent intercept and slope re-estimation models in the Non-Hispanic Black subgroup, respectively. We further found that biased underestimation of unhealthy alcohol use risk persisted across the recalibrated models with regards to both FOR and FNR compared to the original model. There were also minimal changes in FDR and FPR with recalibration.

Figure 3. Observed versus predicted probability of unhealthy alcohol use among Non-Hispanic Black patients in the inpatient cohort according to the (A) original versus intercept re-estimation model, (B) original versus slope re-estimation model, and (C) original versus concurrent intercept and slope re-estimation model. Five bins were used for each graph.

Discussion

Our study demonstrates biased underprediction of unhealthy alcohol use by our NLP classifier for Hispanic patients compared to Non-Hispanic White patients admitted for trauma. For non-trauma hospitalizations, we also demonstrate biased underprediction of unhealthy alcohol use for adult patients under 45 years old compared to adults 45 years and older and Non-Hispanic Black patients compared to Non-Hispanic White patients. Moreover, we show that the greatest improvement in classifier accuracy with recalibration is achieved via concurrent intercept and slope recalibration within the Hispanic subgroup of the trauma cohort. Finally, none of the recalibration methods we implemented adequately addressed the risk underprediction disparities seen across age groups and racial/ethnic identities.

Both of our study cohorts have a median age between 40 and 50 years old and are predominantly male and Non-Hispanic White. The demographic composition of the training cohort is likely responsible for some of the bias seen in our NLP classifier. Our trauma cohort represents all patients who presented to the LUMC Emergency Department for trauma over a span of three years and thus is a better representation of true unhealthy alcohol use prevalence than our inpatient cohort, which includes an oversampling of encounters with a high likelihood of associated unhealthy
alcohol use. Whole-cohort bias metrics are similar across the two cohorts except for the high FDR and FNR in the trauma cohort; these values are likely driven by the lower sensitivity and thus smaller number of TPs in the trauma cohort.

We were able to identify disparities in over- and underprediction of unhealthy alcohol use risk by examining FDR, FPR, FOR, and FNR resulting from our NLP classifier across demographic subgroups. Such algorithmic biases can be introduced in several places along the model development pipeline, including through the training data, model design, and threshold selection. The data used in the development of our NLP classifier resulted from a consecutive sample of patients seen over three years, though approximately half of the eligible patients in this time were not screened for unhealthy alcohol use via the AUDIT. Deviations from true consecutive sampling can result in selection bias due to underrepresentation of certain patient populations; this is particularly problematic when underrepresentation of minority groups is potentially involved. Measurement bias is also a concern when developing NLP tools, especially given existing evidence that implicit racial biases can influence physician language and the knowledge that biased physician language can be perpetuated through NLP features. Biases can further be perpetuated by using a single risk threshold, which can result in a lack of assistive measures for patient populations with systematic risk underprediction or an excess of punitive measures for patient populations with systematic risk overprediction.

Through our attempts to mitigate the biases identified in our unhealthy alcohol use classifier, we found that model recalibration was insufficient to address disparities in risk underprediction-indicating metrics across demographic subgroups. Alternative methods for minimizing bias in classification models can be used, starting with measures to reduce selection and measurement biases during data collection. Additional methods focus specifically on minimizing bias among minority populations through improvements in accuracy, including oversampling with subsequent weighting of minority population data and transfer learning. Our approach of post hoc recalibration within subgroups with biased risk underprediction resulted in some modest improvements in classifier accuracy. However, this method failed to produce improvements in FOR and FNR that would translate to positive effects in clinical practice. Our classifier in its current form stands to recommend assistive unhealthy alcohol use interventions to Hispanic trauma patients less frequently than to Non-Hispanic White trauma patients. Our results likewise demonstrate the value of bias assessment across patient subgroups rather than solely relying on global accuracy metrics of classifier models.

Our study has several limitations. We were unable to compare classifier bias between cohorts due to the use of two different gold standards for identifying unhealthy alcohol use. We were also unable to assess whether using different, sex-dependent AUDIT cutoffs for reference standard unhealthy alcohol use labeling in the trauma cohort introduced or masked classifier bias. Small sample sizes within some subgroups required us to collapse groups down, as in the age groups, or forego analysis altogether, as in the American Indian and Asian racial groups. Counts of FPs, TPs, FN, and TNs were also limited (<25) in four instances. Further, we were unable to assess for bias by gender identity as this data was not available.

We demonstrated the presence of biased unhealthy alcohol use risk underestimation by our NLP classifier among Hispanic compared to Non-Hispanic White trauma inpatients, 18- to 44-year-old compared to 45 years and older medical/surgical inpatients, and Non-Hispanic Black compared to Non-Hispanic White medical/surgical inpatients. We further showed that intercept, slope, and concurrent intercept and slope recalibration resulted in minimal or no improvements in bias-indicating metrics within these subgroups. In summary, we detected bias in our NLP classifier and were unable to adequately address this bias through post hoc recalibration methods. Our results exemplify the importance of integrating bias assessment early into the classifier algorithm development pipeline, ideally in collaboration with health equity researchers and with consideration of the complex nature of bias as it relates to structural health disparities.

References
1. Center for Behavioral Health Statistics and Quality SAaMHSA. Key substance use and mental health indicators in the united states: Results from the 2018 national survey on drug use and health. 2019.
Older adults’ personal health information management: The role and perspective of various healthcare providers

Alyssa L. Bosold, MPH1, Shih-Yin Lin, PhD,2 Jean O. Taylor, PhD1 George Demiris, PhD,3 Anne M. Turner, MD, MLIS, MPH1,4
1University of Washington School of Public Health, Department of Health Services, Seattle, WA, USA; 2New York University Rory Meyers College of Nursing, Hartford Institute for Geriatric Nursing, New York, NY, USA; 3University of Pennsylvania School of Nursing, Department of Biobehavioral Health Sciences, Philadelphia, PA, USA; 4University of Washington School of Medicine, Biomedical Informatics and Medical Education, Seattle, WA, USA

Abstract

The management of personal health information (PHI) by older adults (OAs) takes place within a socio-technical context and requires the support of various stakeholders, including healthcare providers. This study investigates provider roles in supporting OA personal health information management (PHIM), barriers they face, and related design implications for health information technology (HIT). We interviewed 27 providers serving OAs in Seattle, WA. Providers support OA PHIM through medication management, interpreting HI, and providing resources. Barriers to OA PHIM described by providers include (1) challenges with communication between OAs, providers, and caregivers, (2) limited time and resources, and (3) limitations of tools such as secure messaging. Considering these barriers, provider roles, and the socio-technical context for HIT implementation, we recommend the design of HIT that facilitates communication across multiple provider types, integrates caregivers and patient-generated data, supports understanding of OA home environments, and offers credible health resources designed for OAs.

Introduction

The population of older adults (OAs), defined as individuals age 65 and above, is expected to increase from 49.2 million in 2016 to 95 million by 2060 in the United States.1 Aging is associated with greater utilization of healthcare services and more complex health issues,2 resulting in interactions with diverse healthcare providers and an increased amount of personal health information (PHI) to manage.3 Examples of PHI include medical history, tracked symptoms, medication lists, clinic visit summaries, and other information individuals keep about their health.4,5

The work of managing information related to one’s health has been called personal health information management (PHIM), and includes tasks and activities focused on finding, interpreting, using, organizing, and remembering one’s PHI.4,5 This work can help OAs to manage health conditions6 and support healthy aging,7 and is critical for developing systems of self-care, such as medication management. However, PHIM can be challenging for OAs, and is often compounded by conditions commonly seen in OAs, including memory loss or declines in physical health. In contrast to effective PHIM, a lack of accurate and complete information may contribute to poor quality care and health outcomes.8,9 Health information technology (HIT) is increasingly used to support and facilitate PHIM, but HIT design often overlooks the needs of OAs.10,11 Previous studies have demonstrated that utilizing sociotechnical models to examine patient work such as PHIM within organizational, physical, and social contexts can lead to development of HIT that takes into account OA needs.4,12-15

Our prior work from the Studying Older Adults and Researching their Information Needs and Goals (SOARING) project (soaringstudy.org) used a sociotechnical approach to investigate the work system of OA PHIM through extensive interviews with a diverse set of OAs and those involved with OA PHIM.16 Our research, along with others’, confirms that healthcare providers generate, gather, and transfer much of the health information (HI) sought and received by OAs,15,17 and therefore play a unique role in supporting OA PHIM. Further, when we asked OAs and their family and friends about key individuals important to health information management, they named health care providers, primarily physicians, but also pharmacists, social workers, and assisted living (AL) staff.15 As OAs age, others (family and friends and AL staff) provide increasing support for OA PHIM.15,17,19 A depiction of the network by which PHI is transferred in the context of OA PHIM is outlined in Figure 1.
Despite their importance in generating and providing HI, little is known about the role various health care providers play in OA PHIM. Prior research has examined physician and OA information needs in specific situations, such as hospital to home transitions or in shared decision-making. This study looks at a wider ecosystem of providers, seeking input from primary care physicians, gerontologists, medical residents, pharmacists, social workers, and assisted-living staff. We utilized the SEIPS 2.0 model, particularly the components of organization (communication, resources), and technologies and tools, to explore their collective and regular involvement in OA PHIM and the challenges they face, in an effort to help inform the design of HIT supportive of OA PHIM.

![Network for transfer of PHI](image)

**Figure 1.** Network for transfer of PHI

**Methods**

**Recruitment and Participant Characteristics**

We purposively sampled five provider types: primary care providers, specialists (including medical residents and gerontologists), social workers, pharmacists, and assisted living (AL) staff. Participants were initially recruited with emails to assisted living facilities, local hospitals, clinics, geriatric practitioner networks and local senior centers. Following initial recruitment, we recruited additional providers through snowball sampling. We continued to recruit and interview providers within each type until we reached data saturation, outlined by Saunders et al. as hearing similar information from participants. We interviewed 27 providers: eight family doctors and nurse practitioners, seven specialists (five geriatricians and two medical residents), four social workers, four pharmacists, and four AL staff. For the purposes of this paper, we grouped nurse practitioners, family doctors, specialists, and residents together as “medical practitioners,” as themes and roles were similar across these categories. Seventy percent of participants were female, 63% were non-Hispanic white and 96% held an advanced degree (Table 1). The percentage of OAs (>=65yrs) cared for ranged from 15 to 100%.

**Table 1. Participant Characteristics**

<table>
<thead>
<tr>
<th></th>
<th>Whole sample</th>
<th>Medical Practitioner</th>
<th>Social Worker</th>
<th>Pharmacist</th>
<th>Assisted Living Staff</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>27</td>
<td>15</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Age: mean (SD), range</td>
<td>48.1 (13.2), 30-71</td>
<td>48.5 (14.1), 31-71</td>
<td>46.5 (16.9), 32-71</td>
<td>47.5 (9.1), 34-54</td>
<td>48.8 (14.4), 30-65</td>
</tr>
<tr>
<td>Female (%)</td>
<td>70.4</td>
<td>73.3</td>
<td>50</td>
<td>75</td>
<td>75</td>
</tr>
<tr>
<td>Non-Hispanic White (%)</td>
<td>63</td>
<td>60</td>
<td>75</td>
<td>100</td>
<td>75</td>
</tr>
<tr>
<td>Years working with OA: mean (SD), range</td>
<td>15.9(12.0), 1-40.5</td>
<td>15.5 (13.7), 1-40.5</td>
<td>12.3 (6.6), 4-20</td>
<td>13.6 (9.3), 2.5-25</td>
<td>19.8 (12.9), 6-37</td>
</tr>
<tr>
<td>OA in patient panel (%) *</td>
<td>15-100</td>
<td>15-100</td>
<td>60-100</td>
<td>60-90</td>
<td>100</td>
</tr>
</tbody>
</table>
a. Unable to calculate mean and SD due to some participants providing a range, such as “20-30%” or “over 50%”. b. Medical practitioner = family practice/internal medicine physicians, geriatric specialists/fellows, medical residents, and primary care nurse practitioners. c. Social worker = clinical social workers and senior center social workers. d. Pharmacist = geriatric pharmacists, clinical pharmacists, and one pharmacist who coordinated between hospitals and temporary care units and assisted living facilities. e. Assisted living staff = a health services director, an assisted living administrator, a health information manager, and a nursing director.

Data Collection

The SOARING team developed an interview guide and pre-tested it with the team’s clinical advisors before finalizing it. The guide included structured questions about demographics, and a series of semi-structured questions concerning providers’ support of and perspectives on OA PHIM. We conducted in-person one-hour interviews with providers from June 2016 to January 2018. Experienced qualitative researchers (KO and AB) conducted the interviews at providers’ workplaces. Participants received a $50 honorarium. Interview recordings were transcribed verbatim by professional transcriptionists and uploaded to the qualitative online platform Dedoose (Dedoose, Manhattan Beach, CA). The University of Washington’s IRB approved all recruitment materials and procedures (IRB #: STUDY00000084).

Data Analysis

We analyzed interview transcripts using a thematic analysis process. Specifically, after reviewing transcripts, researchers (KO, AT, AB, SL) generated codes, making notes on patterns and insights relevant to OA PHIM. A preliminary codebook was created by one researcher (AB) in conjunction with the research team. Three researchers (KO, AB, SL) then independently applied the codebook to all transcripts and compared their application of the codes, resolving any discrepancies through discussion. After coding transcripts, the team met to discuss emerging themes. Codes fell under three categories: 1) providers’ roles in OA PHIM, 2) PHIM tasks that providers observed among OAs, and 3) the barriers to managing and sharing PHI between OAs and providers. We also examined codes within and across provider types to identify patterns unique to a provider type in each of these three categories. Recognizing parallels between the observations of providers and our prior work with OAs, we analyzed provider responses through the lens of OA PHIM tasks, identifying provider support of, and barriers or facilitators to, each task.

Results

Providers Types, Roles, and HI Exchanged with OAs

Providers described specific tasks they completed in support of OA PHIM, including providing health information (e.g., regarding medications, conditions, or treatments), interpreting or reconciling PHI with OAs and their caregiver, and sharing the OA’s PHI with other providers. Distinctions in the tasks performed by different providers in support of OA PHIM, and variations in the types of information needed, are summarized in Table 2 and detailed below.

Table 2. Summary of Provider Types and the Tasks they Complete to Support OA PHIM

<table>
<thead>
<tr>
<th>Provider Type</th>
<th>Information exchanged</th>
<th>Tasks supporting OA PHIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medical Practitioners</td>
<td>● Exchange a wide range of general medical information, including medical history and medical status</td>
<td>● Interpret medical information and answer OA medical questions</td>
</tr>
<tr>
<td></td>
<td></td>
<td>● Shared decision-making about treatment (interpret information together)</td>
</tr>
<tr>
<td>Pharmacists</td>
<td>● Exchange medication-specific information</td>
<td>● Reconcile medication lists, help with medication management, and develop medication calendars</td>
</tr>
<tr>
<td></td>
<td></td>
<td>● Explain medications and side effects to OAs</td>
</tr>
<tr>
<td></td>
<td></td>
<td>● Confer with medical providers about medication lists</td>
</tr>
<tr>
<td>Social Workers</td>
<td>● Exchange information to gather a holistic picture and address a wide range of social, financial, and emotional needs.</td>
<td>● Provide non-medical resources related to transportation, housing, mental healthcare, legal assistance.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>● Communicate with OA’s physicians.</td>
</tr>
<tr>
<td>Assisted Living Staff</td>
<td>● Information exchanged often depends on the older adults’ level of care. Staff collected information to determine if increased support was necessary.</td>
<td>● Manage all health information for those older adults who need the support</td>
</tr>
<tr>
<td></td>
<td></td>
<td>● Communicate updates on the OA’s health status, or share HI, with family and friends</td>
</tr>
</tbody>
</table>

257
Medical Practitioners

Medical practitioners spoke of documenting visits and collecting information (ex. medical histories, medications) from OAs and other providers. Many discussed the importance of gathering up-to-date histories, contact and medication lists to support the care team as well as OA PHIM. Medical practitioners frequently helped OAs interpret and understand PHI, such as test results and medical instructions, and encouraged OAs to use this information to reach their health goals. Some medical practitioners described the importance of sharing PHI (such as blood pressure logs) to help describe conditions, and facilitate shared decision-making about management strategies with OAs.

Pharmacists

Pharmacists’ support of OA PHIM often involved navigating between OAs and their medical doctors to clarify and update medications or interpret clinic visit summaries. Pharmacists in particular valued up-to-date medication lists:

*The number one thing is an accurate and up-to-date medication list that’s easily accessible to the individual and any key family members or caregivers, as well as something that can be shared easily with the healthcare team.* (P24, pharmacist)

Additionally, all pharmacists described supporting OA PHIM by helping OAs to create and maintain medication lists. For example, three interviewees made calendars to help OAs keep track of medications. One pharmacist even went to OAs’ homes to help reconcile and organize medications.

Social Workers

Social workers addressed a more holistic range of needs, focusing on understanding OAs’ social and emotional needs and connecting OAs with a variety of supportive resources (e.g., transportation, housing, mental healthcare, legal assistance, and power of attorney designation). Social workers had more opportunities to work with OAs in their homes, and communicated with doctors to provide contextual information about OAs’ mental health or home environment. One social worker described how communicating with doctors could help ensure consistency in the information provided to OAs and support adherence to care plans. A doctor similarly described the benefits of delivering messages to OAs through a team-based approach that included social workers.

Assisted Living (AL) Staff

AL staff in our study came from a range of backgrounds and had varied roles. In terms of supporting OA PHIM, AL staff first gathered information centered on admission to assess health status and level of support required:

*It’s always important to get a very clear and accurate picture of the person prior to their moving in…knowing exactly who they are, what their needs are, what their cognitive status is, and their mobility.* (P27, assisted living staff)

Once in the facility, AL staff often monitored and recorded information related to OAs’ health and functional status (e.g., fall incident reports, diagnoses, medication list, contact information of doctors, mobility issues, treatment for current illnesses) to determine needs and provide care. While AL staff often supported independent and collaborative PHIM, unlike other providers, AL staff sometimes took over PHIM tasks for OAs, or worked directly with OAs to collect, record, and track daily PHI. Their role often involved communicating with OAs’ family and friends, requesting information, providing updates about the OAs’ health status and care needs, and sharing records of health care services provided.

Observed OA PHIM

Descriptions of PHIM practices observed by providers fell into three key themes: 1) OAs have diverse PHI organizing and tracking behaviors, 2) OAs have varied medication management practices and 3) OAs seek HI from diverse sources.

Diverse HI organizing and tracking behaviors

Providers reported that OAs demonstrated a wide range of PHI organizing and tracking behaviors, from neatly arranged logs, to relying on memory or the memory of others. Fourteen providers described receiving patient logs from OAs, and had a positive opinion of such patient tracking. One AL staff commented that OAs’ active logging behavior reflected an interest in their health and helped her gauge their need for support.

Varied medication management practices
Twenty-one providers of all types stressed the importance of maintaining accurate and up-to-date medication lists. As with tracking, some OAs had highly organized medication management systems while others did not. Providers described a variety of medication management strategies including written lists, pictures of medications, medication dispensing devices, calling the pharmacist, and having adult children or the AL facility manage their medication. Some providers, however, reported that it was rare to see OAs having an organized medication list, and one commented that lists often did not include drug dosage and frequency.

HI sought from various sources

Twelve providers mentioned that OAs sought and received HI from outside sources including family, friends, the internet, television, advertisements, books, magazines, newspapers, community organizations, and support groups/classes. These observations are generally in line with our prior work exploring OA’s PHIM practices. In addition, our previously documented OA PHIM activities and tasks of seeking, sharing, organizing/tracking, and planning were supported in different ways by healthcare providers in this study. Table 3 highlights how OA activities and tasks were supported, along with some of the related barriers and facilitators to each activity and task. Barriers and facilitators are described in more detail in the section below.

Table 3. Provider Support of OA PHIM Activities

<table>
<thead>
<tr>
<th>Support of OA SEEKING health-related Information (Cont.)</th>
<th>• Communicate with other providers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>o To answer questions about the OA’s health (MP, P, SW, AL)</td>
</tr>
<tr>
<td></td>
<td>o Regarding medication lists, and clarification of correct medications (P)</td>
</tr>
<tr>
<td></td>
<td>o To provide additional context related to OA’s health (SW)</td>
</tr>
<tr>
<td></td>
<td>• Interoperability of systems and time spent gathering HI</td>
</tr>
<tr>
<td></td>
<td>• Lack of communication between providers and incomplete notes</td>
</tr>
<tr>
<td></td>
<td>• Lack of clarity about responsibility for sharing information</td>
</tr>
<tr>
<td></td>
<td>Collaborative provider teams</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Support of OA SHARING - inclusion of others in communication and PHIM</th>
<th>• Communicate with informal caregivers (family and friends) when appropriate (MP, AL)</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Shared decision-making about treatment (MP)</td>
<td>• OA cognitive/physical impairment</td>
</tr>
<tr>
<td>• Develop trusting and supportive relationship with OA and caregivers (SW)</td>
<td>• HIPAA, access to caregiver</td>
</tr>
<tr>
<td>• OA resistance to loss of autonomy</td>
<td></td>
</tr>
<tr>
<td>Collaborative provider teams</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Support of OA TRACKING and ORGANIZING - generating, logging, and organizing health related materials</th>
<th>• Suggest PHIM tools (e.g. logs, charts, pillboxes) (MP, P)</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Reconcile medication lists (P)</td>
<td>• OA cognitive / physical impairment</td>
</tr>
<tr>
<td>• Help OA to organize and manage HI in the home (SW)</td>
<td>• Incomplete provider notes</td>
</tr>
<tr>
<td>• Collect, record, and track daily HI (AL)</td>
<td>• Limited time</td>
</tr>
<tr>
<td>Caregiver present</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Support of OA EMERGENCY PLANNING – preparing and maintaining information in case of a health-related emergency</th>
<th>• Early discussion about care preferences in case of memory loss (MP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>• OA cognitive / physical impairment</td>
<td></td>
</tr>
<tr>
<td>Caregiver present</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Support of OA SEEKING health-related Information</th>
<th>• Communicate with OAs about their health/medical care</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Interpret medical information and answer OA medical questions (MP); explain medications and side effects (P)</td>
<td></td>
</tr>
<tr>
<td>• Provide health information and instructions in a variety of formats (e.g. verbal, written, etc.) (MP)</td>
<td></td>
</tr>
<tr>
<td>• Assist OAs:</td>
<td></td>
</tr>
<tr>
<td>• Look for information to answer health questions and navigate the healthcare system (SW)</td>
<td></td>
</tr>
<tr>
<td>• Discern which information is credible and which medications would be useful vs. not useful (P)</td>
<td></td>
</tr>
<tr>
<td>• OA cognitive / physical impairment</td>
<td></td>
</tr>
<tr>
<td>• Limited time with OA</td>
<td></td>
</tr>
<tr>
<td>• Language barriers</td>
<td></td>
</tr>
<tr>
<td>• HIPAA, access to caregiver</td>
<td></td>
</tr>
</tbody>
</table>

*MP= Medical Practitioner; P= Pharmacist; SW= Social Worker; AL= Assisted Living Staff
Barriers to OA PHIM

We asked providers to identify barriers that they noticed OAs experiencing when practicing PHIM, and challenges to gathering and sharing health information in support of OA PHIM with both OAs and caregivers. Providers recognized a range of OA abilities and involvement with managing PHI, and often described positive experiences, including building trusting, long-term relationships and establishing regular communication electronically or otherwise. We focus on challenges below, but also highlight some of the key facilitators to OA PHIM that emerged from our conversations. Challenges are organized into elements of the socio-technical system that influence OA PHIM, including communication, resources, and PHIM tools.

Communication

Communication between providers and a) OAs, b) informal caregivers, and c) other providers, provided the context for supporting OA PHIM.

Communication with OAs: Nine providers of various types identified cognitive impairments as a barrier to communication. They noted that it was difficult for OAs with cognitive decline to recount medical history or current diagnoses, which made it challenging to determine the accuracy of the information. This hampered effective care-planning and support of OA PHIM. One social worker recommended initiating early conversations about care preferences and care transitions as a pre-emptive measure in case memory loss occurred.

I think nobody likes to have that conversation...about long-term-care planning. But it would be much easier to have a plan in place so that you're not taken by surprise. (P20, social worker)

Four providers described that physical impairments, such as visual or hearing difficulties, presented barriers for accessing and receiving PHI. Consequently, some providers offered information in a variety of formats (e.g., DVDs, phone calls, written information). Three providers, including a pharmacist and two medical practitioners expressed a desire to see OAs' home environments, in order to ensure safe medication management and provide context for management of other conditions.

Communication with Informal Caregivers: In general, providers described caregiver presence during a visit with an OA as helpful for communication and transfer of PHI between the provider and OA. Caregivers provided another ear in handling PHI, remembered things that OAs forgot, and filled in information gaps, particularly if OAs were experiencing cognitive decline.

...having caregivers and family in the room is extremely valuable for the information that they're able to provide that a lot of times the patient cannot. (P2, medical practitioner)

Although caregivers did facilitate communication of PHI, five providers described that it was difficult to contact busy caregivers to obtain and share relevant PHI. In other cases, family tensions created confusion around whether to contact caregivers or which caregiver to contact.

Communication Among Providers: Lack of communication between providers was a major barrier to PHI exchange across provider types. Issues with tools and technologies, such as limited interoperability between different electronic health record (EHR) systems, compounded communication barriers. All providers described spending time tracking down OA PHI by calling or faxing other providers. One medical practitioner described a process of gathering patient records from other healthcare systems, reviewing and entering them into their own systems. The task of gathering information was often done on the provider’s own time. When asked what would improve information exchange, this provider explained:

... [the] ability to get their [OAs'] data from other doctors and what meds they're on, just without having to go through so much rigamarole. I mean...something where everything's on the same system. (P4, medical practitioner)

Two of four pharmacists described not having access to current and complete HI from other facilities, such as hospital discharge notes that were needed to fill prescriptions, or data from other pharmacies to show if prescriptions had already been filled. For example, providers described a lack of clarity about whose responsibility it was to collect and share PHI. Because of these barriers to communication, some providers relied on OAs to bring along updated notes and medication lists to visits, thus increasing the burden for OA PHIM.

In contrast, teams of multiple provider types working collaboratively facilitated communications. One medical practitioner described the advantage of having a pharmacist available to ensure OAs were managing medications
appropriately. Additionally, one medical practitioner explained that having social workers integrated into the healthcare team facilitated communications and provided a more holistic picture of OA health and home environment.

**Resources**

*Time*: All types of providers expressed a desire to have more time with OAs to support PHIM. Several medical practitioners complained that current clinical workflows provide them with limited time to spend with patients. The lack of time limited the ability to collect critical PHI and provide individualized care.

One pharmacist expressed frustration that the system’s focus on the number of filled prescriptions did not allow time for understanding the patient’s needs and delivering quality care. Social workers described OAs who felt frustrated with the fast-paced healthcare system and the lack of human connection.

*Health Information Resources*: A frequently mentioned barrier to effective OA PHIM was the lack of access to plain language materials. Four medical practitioners expressed the need for simplified after-visit instructions and interpretations of lab results, and three pharmacists mentioned simplifying medication inserts so that the risks and benefits were easily understood.

> I think one of the worst things people can have is those darn printouts you get from the pharmacy that give you the gazillion list of possible side effects... nobody looks at them and anybody who does look at them gets terrified. (P23, pharmacist)

In addition, nine providers experienced challenges communicating with OAs who did not speak English, and expressed a need for HI resources in multiple languages. Six providers noted that some OAs had difficulty interpreting and judging the credibility of information gathered online and through other sources, and wanted tools to assist OAs in identifying quality HI sources.

**Tools**

*Patient Portals and Secure Messaging*: Providers were asked about the strengths and challenges of HIT with respect to OA PHIM. Much of their discussion focused on patient portals and secure messaging. Overall, providers felt that patient portals helped to facilitate care and allowed OAs to more easily manage their HI and communicate with providers. However, some providers also described challenges related to secure messaging. Five providers explained that with the introduction of secure messaging, there was an influx of questions from patients, and little time to respond. Four providers described confusion around the use of the portal for urgent messages. As one provider explained, the system had no way to flag messages that described urgent symptoms and should be triaged to a doctor immediately:

> ... I have gotten several symptom-based, “should I go to the emergency room?” type questions...that should have been answered much before they were. So it [the patient portal messaging system] doesn’t always flag urgency. (P14, medical practitioner)

Finally, four providers heard similar complaints from OAs about responses to messaging. OAs were concerned that their messages were not reaching the provider, as there was no confirmation of provider receipt, slow response times, or they received responses from nursing staff instead of their provider.

**Discussion**

We examined the role of different healthcare providers in OA PHIM and identified perceived barriers. Provider roles in supporting OA PHIM necessarily involved the exchange and flow of information, which providers used in unique ways to facilitate optimal OA health. Perceived barriers were best described within the sociotechnical contexts of communication (between OAs, providers, and caregivers), organizational resources (provider time, access to adequate health information, system interoperability), and PHIM related tools (patient portal technologies, secure messaging). Based on our findings, we recommend the following strategies to inform the design of HIT in a way that considers context, addresses barriers, and builds on existing facilitators:

1. **Include non-medical practitioner providers (pharmacists, social workers, assisted living staff) in data sharing agreements or in interoperable medical record systems.**

Consistent with existing literature, our providers identified limited system interoperability and lack of clarity around the responsibilities for transferring and sharing HI as barriers to communicating with other providers.22,27-29 In some cases, this placed more burden on OAs for managing their own PHI and sharing it with different providers. Lack of
access to accurate information between providers not only places additional burden on OAs but can lead to problems with OA safety and poor healthcare outcomes, particularly during transitions in care.29

There is a demonstrated need for improved information sharing between healthcare providers in different settings.22,28 Given the network of providers involved in OA healthcare, data sharing agreements or interoperable EHRs among different provider types could have a positive impact on OA PHIM.30 As supported by our interviews, coordinated approaches between providers such as social workers, pharmacists, and primary care physicians would provide a more holistic picture of OA health, and facilitate consistent information sharing more responsive to OA needs.

HIT that compiles relevant information from various providers would help to streamline the information gathering process, simulating a care team approach, in which multiple providers collaboratively support OA PHIM and healthcare. Time saved in gathering information and filling in information gaps could be invested in developing the patient-doctor relationship, supporting OA PHIM goals, and proactively identifying challenges.

b) **Incorporate options that allow providers to access the environments in which OAs are managing medications and other health issues.**

Another strategy to facilitate a more holistic understanding of OA health is to design HIT that facilitates insight into the context in which OAs manage their health and PHI. Specifically, in line with the desires of pharmacists and medical practitioners in our study, effective PHIM system could accommodate various media (video conferencing, photos, etc.) to allow providers, with the permission of OAs, to view and assess the safety of the home environment, ensure medication management, and identify lifestyle issues.

c) **Facilitate OAs’ understanding of HI through multiple sources and support the interpretation of these sources by trusted providers (creating provider-vetted resources outside of a clinical visit).**

Our findings confirm the difficulties that some OAs have in determining the credibility of online HI resources.15 Incorporating provider-vetted and delivered guides, with a variety of online and offline sources (e.g., request-by-mail articles and brochures) tailored to particular OA needs into HIT would address this need. HI delivered both through these guides, as well as through other mechanisms such as after-visit summaries and lab reports, should be presented in plain language form. In addition, it is important to ensure that information is available in multiple languages and tailored to the needs of OAs and others with visual, physical, or cognitive impairments.

d) **Explore mechanisms to provide formal access to HIT systems for informal caregivers.**

Communication between informal caregivers and providers can provide context for supporting OA PHIM, particularly for those who may experience functional or cognitive limitations over time.11,15 Therefore, it is important to explore a dyadic user experience that will facilitate information sharing and communication between caregivers and providers. However, HIT that supports sharing of HI between multiple stakeholders must take into consideration regulatory and personal barriers related to confidentiality and security.19,31 Allowing OAs to assign permissions and decide what information is shared is critical.

e) **Integrate OA generated data into formal HIT systems.**

As patients are being asked to play a more active role in their own health management, they generate their own data, whether these are logs like those commonly seen in our study, or reflections, preferences, or tracking of physiological variables through wearables or other monitoring devices. These tracking activities become a core aspect of PHIM and should become part of any HIT that is used to store and manage patient data. A HIT system that incorporates OAs’ own tracking and logging of daily behavior also facilitates provider support of PHIM, by allowing providers to generate trends and reports, and offer preventive care measures.30

**Strengths and Limitations**

Examination of multiple provider types and their role in OA PHIM offers a more complex and nuanced picture of OA PHIM. However, we interviewed a small number of providers from one metropolitan area, which may affect generalizability of the results. We did not include provider types such as physical therapists and naturopaths, that may also play a significant role in OA PHIM. While steps were taken to avoid bias, including having multiple researchers participate in coding, personal biases still may have influenced the themes we identified as salient.

**Conclusion**

Our work underscores the importance of the context in which HIT is implemented, including the communication, organizational structures, and tools that influence OA PHIM. Our results showcase the necessity of incorporating
multiple provider types into HIT design to support collaborative, patient-centered, and coordinated care that facilitates PHIM and improves the health of OAs. Although we have focused on HIT solutions, we recognize that our results have broader implications for OA PHIM. Our work revealed barriers in provider workflows, communication, and the allocation of time and resources. More work is needed to further examine barriers and related opportunities for policy and structural solutions to support OA PHIM.

Acknowledgements

This work was supported by funding from the Agency for Healthcare Research and Quality (AHRQ) #R01HS022106. The findings and conclusions expressed here are those of the authors and do not necessarily represent the views of AHRQ. The authors would like to thank the providers who took the time to participate in this study. In addition, we would like to acknowledge Katie Osterhage for her work conducting and analyzing provider interviews.

References


Using Machine Learning to Support Transfer of Best Practices in Healthcare

Sebastian Caldas, M.Sc.¹, Jieshi Chen, M.Sc.¹, Artur Dubrawski, Ph.D.¹
¹Auton Lab, Carnegie Mellon University, Pittsburgh, PA, USA

Abstract
The adoption of best practices has been shown to increase performance in healthcare institutions and is consistently demanded by both patients, payers, and external overseers. Nevertheless, transferring practices between healthcare organizations is a challenging and underexplored task. In this paper, we take a step towards enabling the transfer of best practices by identifying the likely beneficial opportunities for such transfer. Specifically, we analyze the output of machine learning models trained at different organizations with the aims of (i) detecting the opportunity for the transfer of best practices, and (ii) providing a stop-gap solution while the actual transfer process takes place. We show the benefits of this methodology on a dataset of medical inpatient claims, demonstrating our ability to identify practice gaps and to support the transfer processes that address these gaps.

Introduction
When a set of practices an institution has developed over time is known to systematically lead to positive outcomes, it is an enticing target for implementation at other similar organizations. However, endeavors aimed at transferring such best practices can be inefficient and prone to failure, with common reasons being the difficulty to codify tacit knowledge and lack of adequate motivation for the transfer. Just as problematic is the inability to even identify a knowledge or practice gap between organizations, as it prevents the whole transfer process from taking place. In the context of healthcare, transferring best practices can be done to increase the efficiency and effectiveness of health services, and to improve patient outcomes: goals that healthcare organizations are under continuous pressure to pursue. Alas, because of the inherent complexity of healthcare facilities, it is also inherently complex to identify and share best practices among them.

Research into the transfer of best practices in healthcare has been overlooked, and thus the community has a limited understanding of which mechanisms could work best in practice. Previous work has focused on theoretically understanding the transfer mechanisms, mostly borrowing tools from other disciplines. For example, Guzman et al. combined tools from organizational learning and knowledge management with a practice-based perspective, laying down a theoretical framework that aligns the complexity of the practice to be transferred with appropriate transfer strategies. Meanwhile, Berta and Baker described the types of environments and knowledge where a transfer is possible. Finally, Elwyn et al. outlined different reasons that could make the transfer process fail in a hypothetical clinical setting using Szulanski’s sticky knowledge framework.

In this paper, we are interested in enabling the transfer of practices whose outcomes are reflected in the scores of Machine Learning (ML) models, particularly classification models. Given the escalating permeation of Artificial Intelligence in healthcare, examples of such models are numerous and diverse. Some of these include sepsis prediction, vital sign artifact alert adjudication, or the detection of abnormal lengths of hospital stays. In our proposed approach, we rely on multiple organizations having each trained an independent model from their operational data, and on an arbiter or advisory agency regulating the sharing of these models between the organizations.

We focus on the first stage of the transfer process as presented by Elwyn et al., the identification of a possible knowledge or practice gap, and on indirectly supporting the rest of the transfer process (implementation, ramp-up, until integration) by leveraging the knowledge gained from models trained at external organizations. For the former, we use the entropy of an ML model scores as a proxy for the consistency of the underlying practice, as illustrated in Figure 1. For the latter, we make use of novel transfer learning strategies to exploit the available knowledge from the external organizations while the actual best practice is successfully transferred.

We test our methodology on a dataset of medical inpatient claims, where our example ML task is to detect overly long lengths of stays. The proposed approach, however, can be broadly applied to other forms of data, such as Electronic Health Records, and other predictive tasks. In our example, we model selected hospital groups as different organizations, and different diagnosis related groups as different practices. We show that our proposed methodology can identify potential practice gaps that we can later support using transfer learning strategies.
Figure 1. Illustration of our proposed strategy to identify potential practice gaps. Consider two organizations, A and B, each one with its own classification model of some practice: model A and model B, respectively. When we use both models on the data from organization A, the resulting distribution of scores will have different entropies. We use this difference in entropy as a proxy to indicate a potential knowledge or practice gap. In this example, the entropy from model A is greater than the entropy from model B. Thus, we will consider organization B to have a more consistent practice than organization A, and we will recommend a transfer of B’s practice to A.

Motivation

In this section, we present an example to motivate our methodology. In this example, we work backwards from an existing transfer learning solution that proves useful, and arrive at the existence of a practice gap that can explain the apparent utility of the transfer algorithm. In our methodology, explained later in the Methods section, we can identify this gap before applying the transfer learning algorithm. The specific solution we consider was recently proposed by Caldas et al.\textsuperscript{11}. In their algorithm, a clinical organization, when assessing a particular data sample, e.g. a particular patient case, dynamically selects between a locally trained model and a model shared by an external organization. In this section, we focus on presenting and interpreting the results of applying the method on one particular Machine Learning (ML) task.

We consider the task of detecting long lengths of hospital stays in medical inpatient claims. Our claims data comes from the year 2016 Patient Discharge Database (PDD) from California’s Office of Statewide Health Planning and Development (OSHPD), and we solve our task using regularized logistic regression models. We define a long stay as one greater than the national average length of stay for the diagnosis related group (DRG) associated with the claim. We perform the analysis separately for each DRG and for two distinct hospital groups, treating each DRG as a different practice and each hospital group as a different organization. We refer to these organizations as Organization A and B. We provide more specific information about our data and ML models in our Results section.

For this motivating example, we focus on the DRG 291 associated with heart failure and shock with major complications or comorbidity. The results of our suggested dynamic classification algorithm\textsuperscript{11} for this DRG are shown in Table 1. As an example of how to interpret these results, consider Organization B: out of the 860 test cases where the external classifier (from Organization A) is used, 102 cases change predictions from incorrect when using the local classifier, to correct when using the external model. This represents 68% of the total count of hospital stays for which predictions of the two considered models differ, and thus we have enough evidence to discard the hypothesis that such difference is random.

Table 1. Results of Caldas et al.’s\textsuperscript{11} dynamic classification for DRG 291.

<table>
<thead>
<tr>
<th>Organization</th>
<th>Number of test samples</th>
<th>Samples handled better by the external model</th>
<th>Successful flips</th>
<th>% Successful flips</th>
<th>Test p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1118</td>
<td>111</td>
<td>178</td>
<td>41.59%</td>
<td>0.99</td>
</tr>
<tr>
<td>B</td>
<td>1883</td>
<td>860</td>
<td>102</td>
<td>68.00%</td>
<td>3.94 × 10^{-6}</td>
</tr>
</tbody>
</table>
The results in Table 1 indicate an asymmetrical relation between Organizations A and B in terms of knowledge sharing: Organization B observes significant gains from using Organization A’s model, but the reverse is not true. In light of this asymmetry, it is natural to inquire about possible causes of the apparent discrepancy. We explore three hypotheses that could explain it:

1. **Difference in sample sizes**: If Organization A were to have a larger data sample to train its model than Organization B, then we could explore the possibility that the observed asymmetry is due to this difference. However, for DRG 291, Organization B has ~20% more of these claims than Organization A, as seen in Table A.1. in the Appendix. As a result, we discard this hypothesis.

2. **Difference in model performance**: An inherent difference in model performance could explain the asymmetrical relationship observed in Table 1. If this hypothesis were to hold, we would expect the model trained in Organization A to outperform the model trained in B when evaluated on B’s data, and we should observe similar effects when evaluating these models on A’s data. In Figure 2, we see however that the Receiver Operating Characteristic (ROC) curve performances of these models are within each other’s error bands. For a more detailed analysis, we focus on comparing the true negative rates at a fixed 90% true positive rate and construct 95% confidence intervals for the difference in performance between models. Both intervals, which are plotted later on in Figure 5, end up containing 0, further indicating lack of evidence to support the hypothesis that the observed tradeoff in utility of these models can be attributed to difference in their predictive power.

3. **Difference in consistency and quality of care in patient treatment practice**: Having discarded perhaps the most obvious candidate explanations for the observed discrepancy, we turn our attention to the actual practice being modeled. With the available data we cannot directly evaluate consistency of clinical decisions, but we can rely on a proxy: the entropy of the distribution of scores produced by each classifier. We consider a higher entropy to be indicative of a more ambiguous practice in an organization - this happens when a model trained on operational data from this organization yields predictions that are less crisp at discriminating, for example, long and short stays in the hospital (to refer back to our working example). Such differences in the apparent crispness of the decision making process between organizations can exist even if the predictive performances of the models trained to automate such decisions do not differ when measured with common means such as ROC analysis.

We plot the score distributions for both models, evaluated on data from Organization B, in Figure 3. Here, we see that the entropy resulting from model A applied to data in B is indeed lower than the one resulting from applying model B to its own test data. We check that this difference is statistically significant by constructing 95% confidence intervals for the differences in entropy, which we’ll plot later on in Figure 6.

![Figure 2](image-url)  
**Figure 2.** ROC curves for our motivating example. We plot the false positive rate in logarithmic scale for visibility.

Not having insight into actual operations of each organization, we cannot definitely conclude that there indeed exists a knowledge or practice gap by just using medical claims records to support such judgement. But we propose that the presented result could motivate a thorough investigation and, if warranted, a properly designed intervention in organizations that show an opportunity for improving crispness of their practice, unless additional evidence could provide simpler explanations for the results in Table 1.

In our proposed methodology, we are interested in first identifying the apparent knowledge or practice gaps. These gaps will motivate organizational audits and adjustments to existing processes to align performance of the target
organization up to the apparent levels of their peers demonstrating a more consistent, crispier decision-making practice. Such organizational change may be, however, effortless and time consuming to prepare and implement. As a result, we also propose using a transfer learning method\textsuperscript{11} as a stop-gap solution. It will selectively apply the crispier decisions from a better-practice organization when assessing an opportunistically selected subset of all cases for which the external model is expected to confidently outperform the model trained on the local data. This strategy aligns with the transfer process steps outlined by Elwyn et al.,\textsuperscript{4} and it mitigates or at least postpones the effort of implementing a solution that may not be necessary if such algorithmic work-around is acceptable.

![Model A - Entropy: 7.45](image)

![Model B - Entropy: 7.48](image)

**Figure 3.** Score distributions for our motivating example. Model A is more confident in its predictions than model B when evaluating both models at Organization B. This is reflected in a lower entropy for model A.

**Methods**

In this section, we first describe our strategy for identifying possible knowledge or practice gaps between organizations, before sketching the transfer learning technique we recommend to support the transfer process.

*Analyzing model outputs to identify knowledge gaps*

Given a machine learning task, we propose that a well-performing and confident model is indicative of a robust and consistent practice. Consequently, if a model trained in an external organization shows a greater performance or confidence than a locally trained model, this suggests that there is a knowledge or application gap between the institutions. The models should be evaluated on the same set of data for conclusions to be valid. As such, we could imagine that organizations share their trained classifiers, perhaps through an arbiter, and then test and evaluate external models on their own local data. This way, the privacy of the individual organizations’ data can be preserved, a restriction often imposed in the healthcare industry\textsuperscript{12}. For example, in Figure 1, both models, A and B, are evaluated on data from organization A in order to draw conclusions about the state of organization A’s practices.

We outline two possible scenarios that can occur when comparing two models, one local and one external, on data from one organization of interest:

- The case in which one model dominates in terms of some performance metric (for example, Area Under the ROC Curve (ROC AUC), or recall rate at a fixed low false positive rate) is straightforward: if the local model dominates, no adverse gap exists; if the external model dominates, there is a possible knowledge or practice gap. The latter case, however, may not be very common in reality due to the usual degradation of clinical models whenever they are used outside of the institutions they were trained in\textsuperscript{13}.

- The case where the models are similarly well-performing, but show a different degree of confidence in their predictions. In this case, if Organization B produces a more confident model than A, as measured on data from A, then this could be indicative of a more consistent practice implemented in organization B. This is the case we encounter the most often in our experiments. To tackle these situations, we quantify model confidence through the entropy of the distribution of scores generated by the model. This is the same proxy we used in our motivating example to quantify consistency of practice. Note that alternative proxies of quality of practice can also be used instead without changing the proposed framework.

Regardless of the scenario that led to the identification of the potential knowledge or practice gap, it is important to verify that this conclusion is not the product of statistical artifacts. In particular, one should be aware of whether the organizations have different amounts of data to work with. More data is usually associated with better-performing
models, so it is crucial to control for this effect before making any recommendations that will consume resources and impact organizations in the long-term.

*Using transfer learning as a stop-gap solution*

Once an organization has identified a potential local knowledge or practice gap with respect to an external institution, it is natural to look into the adoption of the external practice. However, implementing a new or modifying the existing practice can be a long, time-consuming process. While the structural and organizational efforts towards such adoption are planned and gradually implemented, the local practice is known to be sub-optimal, and thus in detriment to both the institution and some of its patients.

To reap benefits while waiting for the organizational changes to yield actual utility, we suggest using a temporary solution in which the local organization exploits the external model within their current practice. We propose the use of a dynamic classification framework, through which the local organization will select the most appropriate model, local or external, for each new sample. This framework proceeds in two stages:

1. **Estimation of the models’ loss**: Once a new, previously unseen by the model sample arrives, we do not have an associated label for it. To estimate the loss that the local and the external classifiers will incur on it, we make use of the *k*-nearest neighboring samples available in the training data set. Our estimate is the mean of the losses that are incurred on these neighbors. Notice *k* is a hyperparameter we may need to tune via, e.g., cross-validation.

2. **Selection of the most competent model**: Having the loss estimates for the local and the external models, we take the ratio of the estimate for the local classifier over the estimate for the external. This way, *r* will be higher whenever the external classifier seems to be more favorable. Hence, we use the external classifier if *r > r*₀, where the *r*₀ threshold can be optimized on the training or validation set of data to maximize the utility of the framework. In more technical terms, *r*₀ can be chosen to minimize the p-value of a statistical test whose null hypothesis states that using the local classifier on samples where *r > r*₀ provides a better or equal utility than using the external one. We can measure the utility in terms of, e.g., correct predictions.

Another strategy would completely replace the local model with the external one while the practice is being transferred. However, this strategy entails forgoing all the knowledge codified in the local model. It may also not yield any benefits if the performance of both models was judged to be similar from the start. On the other hand, because the proposed dynamic framework is based on the comparative analysis of the local behavior of both models, the local organization can yield benefits from the external model selectively on a subset of their samples, even when global performance metrics of the considered models are comparable.

**Results**

**Data Description**

We use the data from the year 2016 Patient Discharge Database (PDD) from California’s Office of Statewide Health Planning and Development (OSHPD). This dataset consists of over 3.8 million inpatient medical claims, with associated information about the diagnoses and procedures reported in each claim. We limit our study to the top 20 diagnosis related groups (DRGs) with the highest volume of claims, which we model separately as different practices that could be improved. Out of these, we exclude DRGs relevant to newborns to further focus our analysis.

Moreover, for clarity of presentation, we focus on data pertaining to only two hospital groups. Nevertheless, our analyses can be directly applied to larger numbers of organizations, and to entities at different levels of organizational granularity, for instance individual hospitals or practices within a hospital system. Throughout the rest of this text, we refer to these selected hospital groups as Organization A and Organization B. We are left with 14 DRGs representing over 240 thousand inpatient claims. Table A.1. in the Appendix shows descriptions of the considered DRGs as well as sample sizes for each DRG in each organization.

**Machine learning models**

To demonstrate our method, we propose a machine learning task that predicts whether a claim will be associated with an unusually long length of hospital stay. To fulfill this task, we binarize the traditional length of stay prediction task, using the national average length of stay of each DRG as a threshold to determine the labels: a positive label represents a stay longer than the national average while a negative label represents a shorter stay. We do not consider a scenario in which our defined organizations could share their data, replicating the realistic setting in which
institutions are constrained by patient confidentiality and other data privacy considerations\textsuperscript{3}. We end up with one model per DRG, per organization.

We solve the task using regularized logistic regression, but our methodology admits any type of model with an ability to estimate loss of predictions made on individual query data samples. We use 5-way cross-validation to pick the regularization hyperparameter, performing a grid-search over 1,000 settings. We train our models using 70% of available data, and report our results on the remaining 30% held out for testing. We use the true negative rate at a fixed 90% true positive rate (TNR @ 90% TPR) as our evaluation metric. This performance indicator quantifies the desired trade-off between predicting lengths of stay in a normal range while maintaining high sensitivity, but alternative performance metrics can surely be used in its stead as well. We present the performance of the resulting models in Figure 4. We refer to the model trained at Organization A as model A, and similarly for model B.

![Figure 4](image)

**Figure 4.** TNR @ 90% TPR for the trained models. Each model is tested on data from the organization where it was trained. The error bars reflect 95% bootstrap confidence intervals, for which we performed 1,000 resamplings. Note that except for a couple of DRGs, results of these models on their own data are not statistically discernible.

**Identification of knowledge gaps**

We begin by observing the differences in performance between model A and model B once both models are used in the same organization. We plot these differences in Figure 5. We observe that most of our confidence intervals include zero difference, and in these cases we cannot claim the prevalence of one model over another. We see only two intervals that do not contain zero: DRG 392, Organization A and DRG 765, Organization B. Both of these are composed exclusively of positive differences, which indicates that the local model outperforms the external. For these cases, we have enough evidence to discard a possible existence of a knowledge or practice gap, as it is defined in our approach.

Next, we proceed to identify models that are similarly performant on data from a given organization, as seen in Figure 5, but show a difference in entropy of the distribution of their prediction scores. Figure 6 shows these results. Here, we are interested in intervals composed exclusively of positive differences, as they indicate that the entropy of the local model is greater than the entropy of the external one, and a sought-after gap may indeed exist. Table 2 summarizes the DRGs and organizations where we have identified potential knowledge or practice gaps.

**Inspection of statistical artifacts**

As previously mentioned, we need to verify that the potential gaps we identified are not products of statistical artifacts. In particular, we want to make sure that they are not by-products of the differing sample sizes between organizations. To do this, for each DRG, we subsample the training dataset in the organization with the most data until it has the same volume as the other organization. We repeat this random subsampling process ten times and train a completely new model each time. The results of this exercise are shown in Figure 7.

Controlling for size, we see some of our conclusions change: some performance differences we thought were negligible, i.e., the confidence interval contained zero, turn out not to be. This happens for Organization A, DRGs
603, 774 and 775. In these cases, our confidence intervals are composed exclusively of positive differences. We take this as evidence to discard a knowledge gap.

**Figure 5.** Differences in performance when model A and model B are used in the same organization. The presented differences correspond to the performance of the local model minus the performance of the external one. Error bars correspond to bootstrap confidence intervals. We maintain an overall confidence coefficient of 95%, using Bonferroni’s method to correct for making multiple comparisons.

**Figure 6.** Confidence intervals for difference in entropy. The presented differences correspond to the entropy of the local model minus the entropy of the external one. Error bars were drawn in a similar fashion as in Figure 5. We do not plot DRG 392 - Organization A nor DRG 765 - Organization B, as we had already discarded the potential existence of a knowledge gap for these cases due to a difference in model performance.

**Table 2.** Conclusions before inspecting for statistical artifacts. A star (*) indicates a potential knowledge or practice gap, as defined by our framework. A “P” indicates the evidence to discard the potential gap came from the difference in model performance, while an “E” indicates the evidence came from the difference in score entropies.

<table>
<thead>
<tr>
<th>DRG</th>
<th>189</th>
<th>194</th>
<th>291</th>
<th>292</th>
<th>392</th>
<th>470</th>
<th>603</th>
<th>690</th>
<th>765</th>
<th>766</th>
<th>774</th>
<th>775</th>
<th>871</th>
<th>872</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>E</td>
<td>*</td>
<td>E</td>
<td>E</td>
<td>P</td>
<td>E</td>
<td>*</td>
<td>E</td>
<td>E</td>
<td>E</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>B</td>
<td>*</td>
<td>E</td>
<td>*</td>
<td>E</td>
<td>*</td>
<td>*</td>
<td>E</td>
<td>*</td>
<td>P</td>
<td>*</td>
<td>E</td>
<td>E</td>
<td>E</td>
<td>E</td>
</tr>
</tbody>
</table>
Figure 7. Differences in performance and differences in entropy when controlling for sample size. Differences shown correspond to the output of the local model minus the output of the external one. Error bars correspond to confidence intervals where we maintain an overall confidence coefficient of 95% and apply a Bonferroni correction. We construct these intervals by assuming normality in the distribution of the differences. We only plot those DRGs and organizations for which we had identified a potential knowledge or practice gap, as shown in Table 2.

**Evaluation of the proposed transfer learning approach**

We evaluate our transfer learning framework by looking at the samples where the algorithm decides to use the external model, specifically those for which the external model changed the prediction with respect to the local one. We call these flipped samples. If the new prediction of these flipped samples is correct, then we consider it a successful flip. Moreover, if the percentage of flipped samples over all flips is greater than 50%, then we consider the method successful.

Table 3 shows our results for the DRGs and organizations where we identified a knowledge gap. We pay close attention to the columns Train $p$-value and Test $p$-value. They both relate to how confident we are that the percentage of correct flips is lower than 50%, but the latter is only available after evaluation. It follows that if Train $p$-value is too large, we will be discouraged from using this framework at test time, recommending the use of only the local model instead. This happens for three of our identified scenarios: DRG 470, 690 and 766, all for Organization B. As a confidence threshold we used an alpha of 0.05 and corrected for multiple comparisons using Bonferroni’s method. For all other scenarios, we observe both a statistically significant Train $p$-value and Test $p$-value. This means we are both confident about using the method and have enough evidence to claim the method is beneficial on our test data.

**Discussion**

We discuss four axes of our work: our results on the inpatient claims dataset, how this work fits within the broader scope of Machine Learning (ML) techniques, and our limitations and possible future work.

**ML-aided practice transfer for inpatient claims:** Table 3 enumerates the nine DRGs and organizations where our methodology identified a potential practice gap. If these organizations were to undertake transfer processes for these practices, our suggested transfer learning stop-gap solution could support six of these processes, allowing the local organization to reap benefits from the external knowledge, codified through the external model, sooner. In the other three cases (DRG 470, 690 and 766, all for Organization B), our suggested solution recommended using only the local model. All these three cases had a difference in entropy that suggested a more ambiguous local practice, but had some of the lowest absolute differences identified, as observed most clearly in Figure 7. A more robust approach could establish a threshold for the difference in entropy, but this threshold may end up being application dependent.

**New avenues for federated learning:** Federated learning is a novel distributed learning paradigm that allows clinical institutions to collaborate in the training of machine learning models without sharing patient data. The transfer algorithm we leverage was developed in the context of federated learning, exploiting model sharing instead of data sharing between institutions. This work shows a new promising avenue for the application of federated learning: supporting practice transfer processes. This can be done by understanding the contributions the local organization makes to the collaborative model vs. those made by external institutions, as we did in our motivating example.
Table 3. Results of the dynamic classification algorithm we suggest as a stop-gap solution.

<table>
<thead>
<tr>
<th>DRG</th>
<th>Org</th>
<th>Train p-value</th>
<th>Number of test samples</th>
<th>Samples handled better by the external model</th>
<th>Successful flips</th>
<th>% Successful flips</th>
<th>Test p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>189</td>
<td>B</td>
<td>$1.03 \times 10^{-8}$</td>
<td>1561</td>
<td>791</td>
<td>84</td>
<td>78.50%</td>
<td>$1.20 \times 10^{-9}$</td>
</tr>
<tr>
<td>194</td>
<td>A</td>
<td>$1.30 \times 10^{-45}$</td>
<td>595</td>
<td>594</td>
<td>329</td>
<td>81.43%</td>
<td>$2.47 \times 10^{-39}$</td>
</tr>
<tr>
<td>291</td>
<td>B</td>
<td>$3.94 \times 10^{-07}$</td>
<td>1883</td>
<td>860</td>
<td>102</td>
<td>68.00%</td>
<td>$6.15 \times 10^{-06}$</td>
</tr>
<tr>
<td>392</td>
<td>B</td>
<td>$5.61 \times 10^{-05}$</td>
<td>1353</td>
<td>501</td>
<td>54</td>
<td>65.06%</td>
<td>$4.02 \times 10^{-03}$</td>
</tr>
<tr>
<td>470</td>
<td>B</td>
<td>0.86</td>
<td>6305</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>690</td>
<td>B</td>
<td>0.87</td>
<td>527</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>766</td>
<td>B</td>
<td>0.30</td>
<td>3013</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>871</td>
<td>A</td>
<td>$2.42 \times 10^{-67}$</td>
<td>4116</td>
<td>3869</td>
<td>886</td>
<td>73.59%</td>
<td>$1.11 \times 10^{-62}$</td>
</tr>
<tr>
<td>872</td>
<td>A</td>
<td>$7.28 \times 10^{-94}$</td>
<td>1570</td>
<td>962</td>
<td>402</td>
<td>78.51%</td>
<td>$2.17 \times 10^{-40}$</td>
</tr>
</tbody>
</table>

Limitations and future work: There exist other other possible differences between institutions, besides those explored in this work, that can explain apparent practice gaps. Two in particular come to mind: (1) inter-center population heterogeneity, and (2) differing ML capabilities, in terms of the adoption and incorporation of ML practices. Future research efforts should be made to develop and test methods that differentiate between differences in these axes and knowledge gaps.

Future work can also validate the generalizability of the proposed approach by extending it to new multi-source datasets. A more ambitious next step would, ideally, observe the identified practices and validate both the existence of a gap and the utility of the stop-gap solution. Such a research study could be impractical, requiring coordination from multiple institutions and the undertaking of an actual transfer process, which is resource-intensive. Another avenue with a more restricted scope is the identification of more desirable model qualities correlated with best practices attributes that have been identified in theoretical work, beyond high performance and low entropy.

Conclusion

We presented a new approach to identifying opportunities and supporting the transfer of best practices in healthcare using a machine learning methodology tailored to this task. To illustrate this approach, we focused on groups of organizations with practices that can be codified into classification models. By sharing these models and comparing their performance and consistency, we draw conclusions about whether there exist potential discrepancies in the robustness of the underlying practices. If we identify such gaps, we propose using a dynamic classification framework as an immediately available stop-gap solution until the proper organizational investigation and improvement plans to address the apparent gaps are developed and completed.

References


Appendix

Table A.1. Details for each selected DRG.

<table>
<thead>
<tr>
<th>DRG</th>
<th>Description</th>
<th>Number of Claims</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Org. A</td>
</tr>
<tr>
<td>189</td>
<td>Pulmonary Edema &amp; Respiratory Failure</td>
<td>2,989</td>
</tr>
<tr>
<td>194</td>
<td>Simple Pneumonia &amp; Pleurisy W Cc</td>
<td>1,883</td>
</tr>
<tr>
<td>291</td>
<td>Heart Failure &amp; Shock W Mcc</td>
<td>4,624</td>
</tr>
<tr>
<td>292</td>
<td>Heart Failure &amp; Shock W Cc</td>
<td>2,775</td>
</tr>
<tr>
<td>392</td>
<td>Esophagitis, Gastroent &amp; Misc Digest Disorders W/o Mcc</td>
<td>3,837</td>
</tr>
<tr>
<td>470</td>
<td>Major Joint Replacement Or Reattachment Of Lower Extremity W/o Mcc</td>
<td>8,299</td>
</tr>
<tr>
<td>603</td>
<td>Cellulitis W/o Mcc</td>
<td>3,369</td>
</tr>
<tr>
<td>690</td>
<td>Kidney &amp; Urinary Tract Infections W/o Mcc</td>
<td>2,109</td>
</tr>
<tr>
<td>765</td>
<td>Cesarean Section W Cc/mcc</td>
<td>4,387</td>
</tr>
<tr>
<td>766</td>
<td>Cesarean Section W/o Cc/mcc</td>
<td>7,840</td>
</tr>
<tr>
<td>774</td>
<td>Vaginal Delivery W Complicating Diagnoses</td>
<td>3,165</td>
</tr>
<tr>
<td>775</td>
<td>Vaginal Delivery W/o Complicating Diagnoses</td>
<td>22,531</td>
</tr>
<tr>
<td>871</td>
<td>Septicemia Or Severe Sepsis W/o Mv 96+ Hours W Mcc</td>
<td>12,375</td>
</tr>
<tr>
<td>872</td>
<td>Septicemia Or Severe Sepsis W/o Mv 96+ Hours W/o Mcc</td>
<td>4,927</td>
</tr>
</tbody>
</table>
Broken down by bias: Healthcare biases experienced by BIPOC and LGBTQ+ patients

Reggie Casanova-Perez, MS1, Calvin Apodaca1, Emily Bascom1, Deepthi Mohanraj1, Cezanne Lane1, Dhrishi Vidyarthi1, Erin Beneteau, MS, Janice Sabin, PhD, MSW1, Wanda Pratt, PhD1, Nadir Weibel, PhD2, Andrea L. Hartzler, PhD1

1University of Washington, Seattle, Washington; 2University of California San Diego, San Diego, California

Abstract

Bias toward historically marginalized patients affects patient-provider interactions and can lead to lower quality of care and poor health outcomes for patients who are Black, Indigenous, People of Color (BIPOC) and Lesbian, Gay, Bisexual, Transgender and Gender Diverse (LGBTQ+). We gathered experiences with biased healthcare interactions and suggested solutions from 25 BIPOC and LGBTQ+ people. Through qualitative thematic analysis of interviews, we identified ten themes. Eight themes reflect the experience of bias: Transactional Care, Power Inequity, Communication Casualties, Bias-Embedded Medicine, System-level problems, Bigotry in Disguise, Fight or Flight, and The Aftermath. The remaining two themes reflect strategies for improving those experiences: Solutions and Good Experiences. Characterizing these themes and their interconnections is crucial to design effective informatics solutions that can address biases operating in clinical interactions with BIPOC and LGBTQ+ patients, improve the quality of patient-provider interactions, and ultimately promote health equity.

Introduction

Implicit bias shapes healthcare providers' behavior and can result in health disparities based on race, ethnicity, gender, and sexual orientation.1,2,3,4 These biases reflect prejudices and stereotypes that may be subtle and unintentional, and lead to well-recognized inequities in health and healthcare for patients who are Black, Indigenous, People of Color (BIPOC) and Lesbian, Gay, Bisexual, Transgender and Gender Diverse (LGBTQ+).3,4,5 For example, Black patients are prescribed less pain medication,6 receive less aggressive heart attack treatments,7 receive fewer cardiovascular referrals,8 achieve poorer reproductive outcomes,9 experience less patient-centered communication, and rate care quality worse than White patients.9 Similarly, bias towards LGBTQ+ people can result in poor quality of care related to discrimination based on sexual orientation and/or gender identity.10 Avoiding seeking care,11,12 lack of provider knowledge and training about LGBTQ+ health care needs,13 a preference among heterosexual providers towards heterosexual patients,14 and insufficient research about the health of LGBTQ+ populations.15 Implicit biases towards these historically marginalized groups negatively impact patient-provider interactions and produce unnecessary health inequities.

Although many have recognized and attempted to mitigate implicit bias,16 evidence suggests that this bias continues to influence clinical interactions.17 This ongoing problem exposes the need for a deeper understanding of how marginalized and underserved patients experience unfair treatment related to perceived bias and discrimination,18 which can help inform mitigating strategies in the future,19 including informatics solutions. As a first step in the design of automated technology to identify biases operating in patient provider interactions, this qualitative study investigated healthcare biases experienced by BIPOC and LGBTQ+ people and elicited their ideas for improving those experiences in future interactions with healthcare providers. We describe themes that emerged from our qualitative analysis of interviews with BIPOC and LGBTQ+ people, including their experiences of bias, their responses, the consequences, and their ideas for solutions. The research questions this study addresses are:

RQ1. How do BIPOC and LGBTQ+ patients experience biases in healthcare?
RQ2. What solutions do BIPOC and LGBTQ+ patients suggest for improving those experiences in the future?

Methods

Study Design
Data collection included an online survey and a semi-structured interview conducted remotely through Zoom. The University of Washington Institutional Review Board approved this study.
**Participant Recruitment**

Inclusion criteria for participant recruitment were: (1) to be BIPOC and/or LGBTQ+, (2) be 18 years of age or older, and (3) reside in the United States. The research team distributed recruitment flyers through institutional mailing lists, social media, and “community champions” who are patient representatives from BIPOC and LGBTQ+ communities who serve on our project’s advisory board. Those who consented to participate completed the online survey and interview. Recruitment ended when the interviews reached thematic saturation.

**Online survey**

The online survey collected demographics information, including age, gender, race, ethnicity, education, and self-identification as BIPOC and/or LGBTQ+, and experiences of discrimination through the standardized 10-item Day-to-Day Unfair Treatment Scale. This widely used scale collects the frequency of experiencing situations considered to be unfair, such as being treated with less courtesy or respect than other people or being called names, harassed or followed around in stores. Response options were “Never”, “Once”, “2-3 times”, and “4 or more times.” For example, if the respondent answered something other than “Never” to the question “Have you been followed around in stores?”, an additional question asked about the main reason for it with the following options: Ancestry, gender, race, age, religion, height or weight, shade or skin color, sexual orientation, education or income level, physical disability, or other. We did not recruit information about participants’ socio-economic status or insurance coverage.

**Interviews**

Interviews occurred between June and November 2020 via Zoom and lasted one-hour. Except for one interview, two researchers interviewed each participant, one as the primary interviewer following the interview guide and the other one asking clarifying questions. During the first half of the interview, we asked participants to “Tell us about a time when you had an interaction with a doctor where you felt not heard, disrespected, or made uncomfortable?” (RQ1. Experiences). The second half of the interview asked them “If you had the power to change the experience you described, what would you change?” (RQ2. Strategies). After each interview, the researchers made notes summarizing the participants’ experiences and solutions. Interviews were recorded and transcribed for qualitative analysis.

**Data Analysis**

Survey data was summarized with descriptive statistics, and qualitative data was summarized using thematic analysis. Thematic analysis involved developing and refining a codebook and transcript coding. Codebook development started with three researchers (RCP, CL & DV) inductively creating a preliminary codebook with the key themes identified from the initial transcript review. RCP incorporated relevant codes from existing literature and advice from co-authors. Next, two new coders (DM & EB) and an interviewer (CA) familiarized themselves with the data and refined the codebook. The analysis team was composed of four coders (RCP, DM, EB, CA). Two were female (DM & EB), one male (CA), and one non-binary (RCP). Two coders identify as Hispanic/Latino (CA & RCP), one as Asian American (DM) and one as White (EB). The analysis team read two transcripts weekly for codebook refinement and iterated until they reached a consensus. With the refined codebook, the researchers had two sessions of consensus coding where they coded individually two transcripts in Atlas.TI v.9 and manually reviewed the applied codes as a team. After making small adjustments to the codebook for addressing remaining discrepancies, the analysis team coded transcripts in pairs in Atlas.TI until completion.

**Research Context**

In early 2020, the WHO declared a pandemic due to the COVID-19 disease, which forced states and countries to go into lockdown, requiring this study work to take place remotely. COVID-19 brought to light long existing health disparities towards marginalized communities who were disproportionately exposed to the virus because of their ‘essential’ job category which prohibited them from staying home. Additionally, the Black Lives Matter (BLM) movement gained massive support after graphic videos depicting police brutality against Black people went viral on social media. Simultaneously, the American political environment and media was frequently filled with discriminatory comments. These three aspects may have influenced BIPOC or LGBTQ+ people to be more willing to share their negative experiences and how those experiences impacted their life. Collecting the data in these unprecedented circumstances presented a unique opportunity to understand patients’ experiences of bias and discrimination during clinical interactions.

**Results**

**Participants**

Twenty-five participants completed the survey and the interview. Table 1 summarizes their characteristics. Participants ranged in age from 19-60 and were racially diverse. The majority were women, not Hispanic/Latino, and
college educated. Eleven participants were BIPOC, ten reported were BIPOC & LGBTQ+, and three were LGBTQ+. One participant did not report to be BIPOC or LGBTQ+ but described themself as an “older Asian woman.”

**Table 1. Participant Characteristics (n=25)**

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td></td>
</tr>
<tr>
<td>19-29</td>
<td>15 (60%)</td>
</tr>
<tr>
<td>30-50</td>
<td>9 (36%)</td>
</tr>
<tr>
<td>50+</td>
<td>1 (4%)</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
</tr>
<tr>
<td>Woman</td>
<td>16 (64%)</td>
</tr>
<tr>
<td>Man</td>
<td>3 (12%)</td>
</tr>
<tr>
<td>Non-Binary</td>
<td>3 (12%)</td>
</tr>
<tr>
<td>Gender Diverse</td>
<td>3 (12%)</td>
</tr>
<tr>
<td>Race</td>
<td></td>
</tr>
<tr>
<td>Black/African American</td>
<td>5 (20%)</td>
</tr>
<tr>
<td>Chinese</td>
<td>3 (12%)</td>
</tr>
<tr>
<td>Asian Indian</td>
<td>4 (16%)</td>
</tr>
<tr>
<td>Native American/Alaskan Native</td>
<td>1 (4%)</td>
</tr>
<tr>
<td>White</td>
<td>3 (12%)</td>
</tr>
<tr>
<td>Multi-race</td>
<td>6 (24%)</td>
</tr>
<tr>
<td>Other</td>
<td>3 (12%)</td>
</tr>
<tr>
<td>Ethnicity</td>
<td></td>
</tr>
<tr>
<td>Hispanic / Latino/a</td>
<td>7 (28%)</td>
</tr>
<tr>
<td>Not Hispanic / Latino/a</td>
<td>17 (68%)</td>
</tr>
<tr>
<td>Prefer not to disclose</td>
<td>1 (4%)</td>
</tr>
<tr>
<td>Education</td>
<td></td>
</tr>
<tr>
<td>Less than high school</td>
<td>1 (4%)</td>
</tr>
<tr>
<td>High school</td>
<td>1 (4%)</td>
</tr>
<tr>
<td>Some college, no degree</td>
<td>4 (16%)</td>
</tr>
<tr>
<td>Bachelor’s degree</td>
<td>12 (48%)</td>
</tr>
<tr>
<td>Graduate degree</td>
<td>7 (28%)</td>
</tr>
</tbody>
</table>

**Experience of day-to-day unfair treatment**

Table 2 summarizes participants’ responses to the 10-item Day-to-Day Unfair Treatment Scale. Overall, participants reported experiencing substantial levels of unfair treatment. For example, all participants reported receiving poorer service than others and experiencing people acting as if they are superior. Eighty-seven percent of participants reported being treated with less respect and courtesy than others, 83% reported being called names or insulted, and 91% reported experiencing other people ‘acting as if they think you are not smart’. Seventy-five percent of participants said they had been followed around in stores because of racial bias. Of those, 70% reported this was due to race, 10% due to sexual orientation or gender, and the remaining 20% chose other reasons such as height or weight, and shade of skin color.

**Interview themes**

Participants talked about their experiences of bias in the healthcare system, their reaction, and its consequences. They described experiences in different healthcare settings during the entire process for care-seeking, from scheduling the appointment to picking up their medications; explaining why their experiences involved different types of healthcare providers (Physicians, Nurses, Pharmacists). Although participants were not prompted, they organically mentioned healthcare system problems that made them feel not being taken care of and contrasted their negative experiences with positive ones. We identified two types of themes: (A) Experience and (B) Strategy. Experience themes refer to the biased healthcare experiences that contributed to feelings of not being taken care of. We identified and named eight Experience themes: Transactional Care, Power Inequity, Communication Casualties, Bias-Embedded Medicine, System-level problems, Bigotry in Disguise, Fight or Flight, The Aftermath. We also report on interconnections among these themes. Strategy themes refer to insights for mitigating experiences of bias in healthcare. We identified two Strategy themes: Good Experiences and Solutions.

**Table 2. Responses to the Day-to-Day Unfair Treatment Scale (n=24)**

<table>
<thead>
<tr>
<th>In your day-to-day life, how often do any of the following things happen to you?</th>
<th>Never</th>
<th>1 Time</th>
<th>2 - 3 Times</th>
<th>4+ Times</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. You have been treated with less courtesy than other people are.</td>
<td>12.50%</td>
<td>12.50%</td>
<td>33%</td>
<td>42%</td>
</tr>
<tr>
<td>2. You have been treated with less respect than other people are.</td>
<td>12.50%</td>
<td>8%</td>
<td>42%</td>
<td>37.50%</td>
</tr>
<tr>
<td>3. You have received poorer service than other people at restaurants or stores.</td>
<td>0%</td>
<td>25%</td>
<td>50%</td>
<td>25%</td>
</tr>
<tr>
<td>4. People have acted as if they think you are not smart.</td>
<td>9%</td>
<td>25%</td>
<td>33%</td>
<td>33%</td>
</tr>
<tr>
<td>5. People have acted as if they are afraid of you.</td>
<td>46%</td>
<td>25%</td>
<td>8%</td>
<td>21%</td>
</tr>
<tr>
<td>6. People have acted as if they think you are dishonest.</td>
<td>33%</td>
<td>21%</td>
<td>29%</td>
<td>17%</td>
</tr>
<tr>
<td>7. People have acted as if they’re better than you are.</td>
<td>0%</td>
<td>17%</td>
<td>29%</td>
<td>54%</td>
</tr>
<tr>
<td>8. You have been called names or insulted.</td>
<td>17%</td>
<td>21%</td>
<td>29%</td>
<td>33%</td>
</tr>
<tr>
<td>9. You have been threatened or harassed.</td>
<td>21%</td>
<td>29%</td>
<td>17%</td>
<td>33%</td>
</tr>
<tr>
<td>10. You have been followed around in stores.</td>
<td>25%</td>
<td>21%</td>
<td>25%</td>
<td>29%</td>
</tr>
</tbody>
</table>
A. Experience with bias in healthcare

1. Transactional care

Transactional care refers to a provider treating an appointment as a job rather than an opportunity to look after the patient’s well-being. This theme refers to the patients’ experience of the appointment. Participants told us about incidents regarding providers rushing the appointment, not being thorough enough or only treating symptoms, refusing to take care of more than one medical concern in one appointment, and being uninformative with the patient about their medical reasoning, diagnoses, treatments, or alternatives. Participants expressed feeling unseen in transactions:

“I talked about being at the hospital and feeling I was just like, an assignment or some sort... of just one thing to do, like, 'okay, this is what you have to check the vitals, and this is where you need to give him this specific thing.' But not really being felt like I was really being seen.” – PT18, BIPOC & LGBTQ+ Non-binary.

“That's what I always hear from my previous PCP is, 'Oh, we'll see if it gets better or take some over-the-counter medication for this' or, 'Oh, you know, just change your diet or exercise more.' But I didn't feel like he was really paying attention to, you know, what I actually was going through.” – PT01, BIPOC & LGBTQ+ Man

2. Power Inequity

Power Inequity refers to the power imbalance between patients and providers that affected rapport building. This imbalance was often described as due to a knowledge difference between patients and providers. This imbalance made some participants feel that the provider judged them not to be smart enough to understand medical information, identify valid medical concerns, or make the best medical decisions. This judgment made some participants feel powerless and have no say in their own health or how they want to manage it. For example, one participant told us:

“I do think that it also comes with that power dynamic of like having more control as a doctor. As a professional you can say like 'I have all these degrees and have all these years of experience, and you're just a patient’ ... you don't have as much power in the situation, so they can kind of have control.” – PT08, BIPOC & LGBTQ+ Woman.

Incidents for this theme involve providers making all the medical decisions without including the patient in the process, not building rapport with the patient, dismissing the patient without giving them an explanation, and making the patient feel that they do not have the qualifications for making medical decisions. For example, PT13 felt dismissed:

“I was asking the doctor what he thought would be like the cause of my symptoms. I suggested something, and he just kind of brushed it off. I feel like I wish he would have just explained why that wouldn't have been the case, rather than brushing it off. It kind of left me wondering why did he brush it off? was it a dumb reason or something?” – PT13, BIPOC Woman.

3. Communication Casualties

Communication Casualties refers to verbal or non-verbal cues that the patient identified as uncomfortable, awkward, or inappropriate. These negative cues made the patient feel unwelcomed, judged, or not taken care of. Participants described how some providers’ comments negatively impacted their ability to receive care. Negative verbal cues refer to providers saying inappropriate comments to the patient, scolding them, or verbally judging their actions. Non-verbal cues refer to behaviors that made the patient feel that the physician was not paying attention to them, wanted the appointment to be over, or is did not understand their needs; these nonverbal cues included lack of eye contact, approaching the door when the patient was speaking, or having a judgmental tone of voice. For example, PT15 described negative cues expressed by a healthcare provider:

“I was seated, she was standing above me hands on her hips, literally lecturing me as if she was my mother. And flat out said 'if you were my son, this is what I would say to you’” – PT15, BIPOC & LGBTQ+ Man.

Several participants talked about the provider transmitting a ‘vibe’ that made them not trust the providers even if they could not explain the communication cue explicitly in words. PT14 describes the “vibe” of a negative interaction:

“I told (the provider) about my issue and (they) don't really offer any suggestions, and it was so tough. You can't explain vibes, right? There's no definition, it just felt awkward and weird” – PT14, BIPOC & LGBTQ+ Woman.

4. Bias-Embedded Medicine

Bias-Embedded Medicine refers to instances in which the patient felt they were treated unfairly because of their personal characteristics, such as race, ethnicity, gender identity, sexual orientation, or physical characteristics (e.g., weight). Because of these possible assumptions, patients had a hard time accessing tests, treatment, or obtaining a diagnosis. Incidents for this theme are misunderstanding of pain, BIPOC & LGBTQ+ stereotyping, and other types of
situations, such as weight or age. Misunderstanding of pain was when providers were thought to assume that the patient (BIPOC patients usually) was exaggerating their pain to get pain medication. For example, PT21 told us:

“I said, ‘Could you please do that [MRI] because I’m a movement performance artist. So, I would jeopardize my entire well-being, career, and financial status if I injured myself further. It's really, really important I treat myself for the exact condition.’ He said, well, ‘we are deeming it medically unnecessary because we do not feel it’s necessary for you to have an MRI because we feel this information is good enough’” – PT21, BIPOC Woman.

BIPOC & LGBTQ+ stereotyping is when the provider adopts certain behaviors and characteristics after making assumptions related to the patient’s demographic. Other types of stereotyping included assumptions based on reasons, like age or weight. For example, one participant described an experience of stereotyping:

“I saw written down ‘high-risk homosexual behavior’. She asked me if I ever had sex with men. She didn’t ask me the last time I had sex with men, and it couldn’t have been because I hadn’t had sex in like months, (…) I just dislike for her to make that assumption, that I was out there having unprotected sex” – PT23, BIPOC & LGBTQ+ Non-binary.

5. System-level problems
System-level problems refer to the patient feeling that the healthcare system is unwelcoming and difficult to navigate. Incidents in this theme related to the medical record, such as trouble accessing their records, or problems like confusing insurance systems, non-inclusive medical forms, and the lack of diversity in the medical workforce. For example, one participant described inaccurate medical records:

“After discharging me, (the physician) wrote in the chart notes that I had refused care. And I said, ‘That’s not true, this is what he told me, and I wasn’t going to stay there for things to get worse’” – PT06, BIPOC & LGBTQ+ Woman.

Another participant describes the lack of diversity they perceive in the healthcare workforce:

“There’s a real issue with supporting doctors who look like their patients. The only people of color are the ones cleaning, or in the cafeteria. I think diversity is such a huge part of it... I’m not saying that all doctors need to be people of color but, maybe seeing more people who just aren’t in that standard mold that we’ve created healthcare to follow.” - PT14, BIPOC & LGBTQ+ Woman.

6. Bigotry in Disguise
Bigotry in Disguise refers to when participants perceived providers to have implicit discriminatory attitudes or behaviors towards BIPOC and LGBTQ+ patients. Examples are the provider not knowing how to talk or take care of gender-diverse patients or appearing unaware of important needs for BIPOC patients, such as the prevalence of specific conditions among certain races/ethnicities. Participants understood that providers are not supposed to know everything but wanted providers to be honest about being unknowable. For example, PT21 describes perceptions of a provider’s disregard for her South Asian ancestry:

“I said, ‘There’s a British study done on South Asians and cardiovascular disease and metabolic syndrome. There should be specifications as to what is healthy (for South Asians)’. He dismissed me and said ‘all that matters is you do x, y, and z, which is not only for South Asians, just start, just do this’. And I said, ‘okay, you don’t know about the South Asian Studies’. And he said, ‘I’ve never heard of any of this, and it doesn’t matter’” – PT21, BIPOC Woman.

Another incident is when patients feel ‘othered’ by providers, meaning that the provider makes a rigid distinction between the provider’s and the patient’s community. These distinctions peaked when BIPOC participants saw providers treating White patients better than them (i.e., racial favoritism). These discriminatory attitudes were also experienced by LGBTQ+ participants. For example, PT25 describes a provider’s bias toward their girlfriend:

“I said, “My partner has a penis. Does that affect this or that?” And he laughed and said, “Oh, well, I’m sure your boyfriend, yada, yada, yada.” It made me so angry, because not only was he… I guess, for my body, I can justify it, and I can understand not everybody is super educated about trans issues or pronouns, whatever. I can excuse that even if it’s not right, to help me get through the process and minimize my discomfort, but as soon as he disrespected my girlfriend, I got mad, and I just stopped talking, started giving very short, to-the-point answers, because I didn’t wanna get angry and I wasn’t going to explain everything in a hospital gown.” – PT25, LGBTQ+ Multi-gender.

7. Fight or Flight
Fight or Flight refers to responses in which the patient realized they were being treated unfairly. There were two distinctly different types of responses: Active and Passive. The active response is when the patient self-advocates and demands to be treated with respect (i.e., fight). The passive response is when the patient says nothing and waits until the interaction is over (i.e., flight) for fear of repercussions. For example, PT21 described her “fight” response:
"I was so angry, and I told him 'I'm extremely upset with how you treated me, the difference in statistics for a South Asian person and a white person are very, very different. You're dismissing me. You're just dismissing me because I'm young and I'm brown, and you don't think my statistics are bad enough, and you've already told me that you think I'm fat, so I'm not going to see you again.'" – PT21, BIPOC Woman.

Surprisingly, several participants, independent of whether they had an active or passive response, justified the providers’ behavior for treating them unfairly. For example, PT04 told us:

“Okay, giving them the benefit of the doubt, they were probably just following protocol. That's where I think that maybe healthcare procedures need to account for looking more like situation alliance.” – PT04, BIPOC Woman.

8. The Aftermath

The Aftermath refers to the consequences that biased experiences had in patients’ lives. The most serious was the health-related consequences, where a patient’s conditions got worse following a provider’s dismissal. Other serious consequences were the patient self-medicating, delayed or avoided healthcare unless it was an emergency, or simply mistrusted the healthcare system altogether. For example, PT11 describes consequences of such dismissal:

"I was overweight but active, I complained about knee pain all the time, and doctors always said 'lose weight.' Well, one time on vacation, I got injured and had to go get X-rays done on my knee. When the X-rays came back, the doctor was, like, “you have arthritis, and also your kneecap is misaligned.” My first reaction was 'I'm not even 30, and I have arthritis?' and then it was anger because I have been dismissed for like a decade". - PT11, BIPOC & LGBTQ+.

Other consequences were changing providers searching for one who would treat them fairly, or start having “covering” behaviors, where the patient would change their physical presentation in the hope that providers will treat them better:

"This is another thing too that I have to do - I'll make sure that I'm suited that day. Like head to toe. Because even that sometimes can help the experience. Let me put on my professional attire. Sadly, it still doesn't work... I actually have to prepare myself in that way and be selective about what I choose to wear that day. Like I wouldn't want to go in like my leggings." – PT16, BIPOC Woman. Quote about covering behaviors

Interconnections among themes related to participants’ experiences with bias in healthcare

Identified themes are intrinsically connected. Figure 1 is a visual representation of interconnection among themes within three nested dimensions: the healthcare system (purple rectangle), the appointment (yellow rectangle), and the patient-provider interaction (green). The eight Experience themes are in bold. The first dimension, the healthcare system, includes the ‘System-level problems’ since participants mentioned structural problems as a challenge when visiting the provider. In the appointment dimension, we found ‘Transactional Care’ since it refers to problems during appointments. In the appointment dimension, we found two entities (in blue), the provider and the patient, connected within the patient-provider interaction dimension. The provider can bring the ‘Bias-Embedded Medicine’ and ‘Bigotry in Disguise’ into the interaction which the patient can perceive as ‘Power Inequity’ and ‘Communication Casualties,’ leaving the decision of “Fight or Flight” to the patient. After the appointment, ‘The Aftermath’ could result in the patient choosing to leave the healthcare system altogether, so it sits halfway outside the healthcare system dimension.

![Figure 1. Interconnections among themes related to participants’ experiences of bias in healthcare](image-url)
B. Strategies to mitigate the experience of bias in healthcare

The second half of the interview focused on the participants’ suggested solutions about how to prevent these biased healthcare experiences from repeating. Many participants organically contrasted their negative experiences with good experiences. We report both participant-generated solutions along with those good experiences since they can display well-received providers’ actions and could help guide mitigating strategies.

1. Solutions

Solutions refer to the participants’ ideas about what could help prevent the biased interactions from repeating. Participants’ ideas include having patient advocates accompany them to feel more comfortable at the providers’ office, providing feedback to the provider to let them know what went wrong during the interaction, enhance provider training on interpersonal skills and cultural competence towards BIPOC and LGBTQ+ patients, and promote diversity in the healthcare workforce. For example, PT16 describes providing feedback:

“The goal, hopefully, from the doctor's perspective, would be to have this honest feedback that they could grow from and implement changes for. It's tough, but you know it's a necessity.” – PT16, BIPOC Woman.

Participants mostly referred to non-technological solutions, but some mentioned the use of technology to address the solutions. For example, PT02 proposed using technology for patient advocacy:

“I think it would be useful in the doctor's office if there was a robot that could be like 'wait, doctor, let the patient talk' because I feel like doctors are always in a rush” – PT02, BIPOC Woman.

2. Good experiences with healthcare

When talking about negative experiences with providers, most participants organically compared those experiences with good experiences they had with other providers. For BIPOC patients, good experiences were the ones when the provider did not doubt their symptoms. For LGBTQ+ patients, good experiences happened when the provider validated their identity and did not make assumptions about them or their partner. For both groups, good experiences were when the provider treated them with respect, carefully listened, took their needs into consideration, gave a thorough explanation of the diagnosis and treatment, genuinely worked in rapport building, and were noticeably interested in reducing the power differences (e.g., sitting at the patients’ eye level, looked them in the eyes). For example, PT22 describe feeling respected by the provider:

“I think it was that my first doctor, he respected that I wouldn't have been there if I hadn't done a ton of research, and he also respected the fact that I had been working in the medical field for 11 years. That I wasn't just an internet researcher. So, he wasn't wasting time trying to dumb things down, and he explained what needed to happen and why it needed to happen” – PT22, LGBTQ Transgender Woman

PT16 describes feeling a warm welcome from a provider:

“He spent time with us, like he wanted to know what I was doing for my well-being, and I talked to him about my arthritis and using CBD, and I went there for my lungs (laugh). And that was a cool experience. And my daughter had just turned 18, so we were in separate rooms, and he was 'if you guys are cool with it, we can sit together.' It was really warm, and this is a walk-in clinic.” – PT16, BIPOC Woman

Discussion

As a critical first step to inform design of automated technology to identify how biases operate in patient provider clinical interactions, we interviewed 25 BIPOC and/or LGBTQ+ participants about biased healthcare experiences. We found eight themes related to the experiences and two themes about the strategies to mitigate bias in the future. Experience themes were Transactional Care, Power Inequity, Communication Casualties, Bias-Embedded Medicine, System-level problems, Bigotry in Disguise, Fight or Flight, and The Aftermath. We presented a visual representation on the interconnections among the eight experience themes. Strategy themes were Solutions and Good Experiences. The ten themes presented in this study contribute to the understanding of biased healthcare experiences from the perspective of BIPOC and LGBTQ+ patients. Although prior research examines impacts of implicit bias on healthcare (1)(3)(4) we are contributing to filling the gap that exists surrounding how to apply the perspectives of BIPOC and LGBTQ+ patients on biased healthcare experiences and gathers their ideas for designing informatics solutions. Health informatics researchers and practitioners can leverage our findings to inform the development of technological innovations that improve the quality of patient-provider interaction by addressing implicit bias.

The experiences we heard from both groups support prior work related to not being heard22 and being dismissed by providers.23 BIPOC participants experiences related to racial discrimination in healthcare,24 measurements of
perceived racial/ethnic discrimination, its impact on health and well-being, physical and mental health outcomes, the perpetuation of health inequities, and its impact to patient-provider communication. Experiences from LGBTQ+ participants are consistent with literature reporting poor quality of care due to discrimination based on sexual orientation or gender identity, lack of provider knowledge about LGBTQ+ health care needs, and a preference towards heterosexual patients. We corroborated that institutional racism and homo-/transphobia remain problems in healthcare. Themes related to the patients’ response and consequences from a biased interaction, support prior work about medical mistrust originating from perceived discrimination experiences. This understanding of the “Fight or Flight” and “The Aftermath” themes can inform a better healthcare service design for improving patients’ experiences in the healthcare system. These themes and their interconnection allowed us to understand how BIPOC and LGBTQ+ patients experience biases in healthcare, our RQ1.

The “Good Experiences” that participants shared demonstrated that some healthcare providers may have noted these problems and contributed to enhancing the experiences of patients from marginalized groups. Simple non-technical solutions might go a long way to help patients feel validated and cared for, such as a questionnaire for the patient to fill out their sexual orientation and gender identity. Thus, health informatics researchers and practitioners can leverage simple and familiar technologies (e.g., electronic questionnaires) to improve patient-provider interactions. Solutions that aim to enhance providers’ training reflect the need for medical education curriculum with a strong, fully integrated, culturally competent component and content on the medical needs of the BIPOC and LGBTQ+ community. Diversifying the healthcare workforce is another priority in the quest for more equitable care and strategic use of technology in healthcare to achieve health equity (i.e., “TechQuity”). The themes we identified allow us to understand how BIPOC and LGBTQ+ patients want to improve healthcare experiences, which provides critical guidance for all of these potential solutions.

This study has a number of limitations and strengths. One significant limitation is sampling bias. Due to COVID-19 restrictions, we were only able to recruit through virtual means which skewed our sample to be younger and college educated, and potentially socioeconomically advantaged. Our sample was likely limited due to the technological requirements for participating in the interview (e.g., device compatible with Zoom software with a microphone, stable internet connection), which could have prevented people with non-technical backgrounds from participating. Another limitation is recall bias, which could have impacted the experiences the participants reported. Our use of recently minted terminology such as “BIPOC” could also have had an impact on the age of participants because this term may have been more familiar to younger people. Finally, we only interviewed three participants who were LGBTQ+ only, which limits the transferability of findings for this population. However, we captured perspectives from ten participants who were both BIPOC and LGBTQ+, which offered important insights on intersectional experiences.

We demonstrate that there is an opportunity for the biomedical informatics community to establish a more collaborative environment between patients and informaticians for solutions on how to improve the patient-provider interaction, provide better patient-centered care, and incorporate marginalized patients’ needs. Patients’ perspectives about biased healthcare experiences have laid the groundwork for using these patient reported themes in a range of future work to promote TechQuity. In particular, we are using findings to inform co-design work with patients and providers inform automated assessment tools that raise awareness of hidden bias in patient-provider communication. Future work should also examine experiences of intersectional individuals in greater depth and how this intersectionality can be represented and incorporated into technological solutions.

Conclusion

We interviewed 25 BIPOC and/or LGBTQ+ participants who described biased healthcare experiences when visiting a healthcare provider. Through qualitative thematic analysis, we describe ten themes that describe their experiences and suggested solutions to mitigate those experiences in the future. Characterizing these themes and their interconnections lays a foundation for understanding lived experiences of patients from marginalized groups that can critically guide the design of informatics solutions to mitigate those biases, improve the quality of patient-provider interactions, and ultimately promote health equity.

Acknowledgements

The study is supported by #1R01LM013301. We want to acknowledge our participants for sharing their experiences, community champions who helped us recruit, and our advisory board members who helped frame the study.
References


Unpacking the Drop in COVID-19 Case Fatality Rates:
A Study of National and Florida Line-Level Data

Cheng Cheng*, B.S., Helen Zhou*, M.S., M.Eng.,
Jeremy C. Weiss, M.D., Ph.D., Zachary C. Lipton, Ph.D.
Carnegie Mellon University, Pittsburgh, PA, USA

Abstract  Since the COVID-19 pandemic began, the United States’s case fatality rate (CFR) has plummeted. Using national and Florida data, we unpack the drop in CFR between April and December 2020, accounting for such confounders as expanded testing, age distribution shift, and detection-to-death lags. Guided by the insight that treatment improvements in this period should correspond to decreases in hospitalization fatality rate (HFR), and using a block-bootstrapping procedure to quantify uncertainty, we find that although treatment improvements do not follow the same trajectory in Florida and nationally (with Florida undergoing a comparatively severe second peak), by December, significant improvements are observed both in Florida and nationally (at least 17% and 55% respectively). These estimates paint a more realistic picture of improvements than the drop in aggregate CFR (70.8%–91.1%). We publish a website where users can apply our analyses to selected demographics, regions, and dates of interest.

Introduction
Over the past year, the coronavirus (COVID-19) pandemic has continually evolved, with disease outbreaks expanding and contracting; lockdown measures tightening and loosening; testing capacity (mostly) increasing; and treatments protocols evolving. Decision-makers trying to maintain a grasp of the rapidly unfolding situation face a few key questions (among others): Have new treatment protocols improved outcomes over time? Is COVID-19 fatality decreasing? How does the actual infection rate compare to at previous dates? To what extent are rising case counts artifacts of expanded testing?

The most widely reported statistics for monitoring the pandemic are confirmed cases and deaths. At first glance, the two appear to tell divergent stories about the trajectory of the pandemic over time. Confirmed cases peak for the first time in April (with a 7-day average of nearly 32,000 daily cases), peak again with more than twice as many cases in a second wave in July (nearly 67,000), and yet again in a much larger third wave in December (nearly 230,000) (Figure 1, left panel). However, reported deaths appear to tell a contradictory story concerning the relative severity of the three waves (Figure 1, middle panel). Dividing deaths by cases, we observe that the reported case fatality rate has fallen dramatically after the first wave, from nearly 7.9% at the height of the first wave in mid-April to the 0.7%–2.3% range since July. (Figure 1, right panel). There are several plausible explanations for this decline, each with significant policy implications. Thus far, academic and public discourse has centered around the following hypotheses:

(H1) The age distribution of infected patients has shifted, altering the CFR due to the higher risk among the elderly.1–4
(H2) Increases in testing capacity have driven down the CFR due to a rising number of tests catching milder cases.5–7
(H3) Apparent shifts in case fatality rate, are artifacts due to the delay between detection and fatality.1,7,8
(H4) Treatment has improved as doctors grow more experienced and new therapeutics become available.3,9–12
(H5) The disease itself is mutating, leading to changes in the actual infection fatality rate over time.5,13
(H6) Social distancing has reduced the viral load that individuals are exposed to, resulting in milder infections.13–15

Note that H1–H3 can be misleading if not sufficiently accounted for. Due to Simpson’s paradox, since there are large differences in fatality between different age groups, if the age distribution shifts (H1) we could easily observe an overall decrease in CFR despite increasing CFRs in every age group (or vice versa). Additionally, testing ramp-up (H2) and delays between detection and fatality (H3) can cause the behavior of case fatality rate to diverge substantially from the behavior of the true infection fatality rate. Thus, CFR can be a poor proxy for actual infection fatality rate (IFR).

On the other hand, the last three phenomena—improved treatments, disease mutation, and changing viral load—correspond to actual reductions in mortality and could be grounds for policy changes. This work demonstrates how

*Equal contribution.
While these clinical trials have evaluated the effects of specific treatments in their identified target populations, our work studies the broader impacts of treatment improvements over time at a larger national scale. As far as we are aware, this is the largest national-scale (588,126 hospitalizations, 10.3 million cases) data-driven visualization, and analytical code pipeline for both the CDC and Florida COVID-19 line-level public datasets.

In particular, we argue that complete and accurate age-stratified, line-level hospitalization data is pivotal for distinguishing true improvements from artifacts. Hospitalizations should be less influenced by testing capacity than confirmed cases, and less influenced by treatment efficacy than deaths. While there may have been changes in admitting criteria at the very worst moments, for example, when New York hospital demand exceeded capacity in late March, for the most part, criteria for inpatient hospitalization is relatively consistent across time periods. Additionally, in the study time period (April 1st to December 1st), treatment improvements mostly targeted hospitalized COVID-19 patients.

Importantly, this analysis yields several important observations: (i) large increases in testing do occur between the waves but do not explain them away; (ii) since age distributions shifted substantially between the first and second waves, age must be accounted for in order to separate out the effects of treatment from age shift; (iii) between the first and second waves age-stratified HFRs improved substantially in the national data (with HFR decreasing by as little as 27% in the 80+ age group and as much as 37% in the 30-39 age group), but were relatively unchanged in Florida (with a slight increase in HFR by as little as 2.9% in the 80+ age group and as much as 13% in the 60-69 age group); (iv) by December 1st, both Florida and national data suggest significant decreases in HFR since April 1st—at least 17% in Florida and at least 55% nationally in every age group; and (v) comprehensive age-stratified hospitalization data is of central importance to providing situational awareness during the COVID-19 pandemic.

As far as we are aware, this is the largest national-scale (588,126 hospitalizations, 10.3 million cases) data-driven analysis to quantify and account for all three artifacts (age distribution shift, increased testing, and detection-to-fatality delay) when estimating treatment improvements. To allow users to apply our analyses to time ranges, states, and demographics of interest, we publicly release an interactive web application. Finally, we release a full pre-processing, visualization, and analytical code pipeline for both the CDC and Florida COVID-19 line-level public datasets.

Related Work
Several COVID-19 treatments were developed over the study time range (April 1st to December 1st), and each underwent randomized controlled trials testing for its individual efficacy. Dexamethosone resulted in a lower 28-day mortality when estimating treatment improvements. To allow users to apply our analyses to time ranges, states, and demographics of interest, we publicly release an interactive web application. Finally, we release a full pre-processing, visualization, and analytical code pipeline for both the CDC and Florida COVID-19 line-level public datasets.
To get a holistic sense of improvements over time several studies have examined CFRs. In a study of 53 countries, all but ten were found to have lower CFRs in the second wave compared to the first. However, as delineated in our introduction, confounding factors such as shifting age distribution (H1), testing capacity (H2), and detection-to-death lags (H3) can lead to misleading interpretations of the CFR. For example, when comparing CFR by age group in Italy and China, Onder et. al. suggested variation in testing strategies as a possible explanation for discrepancies. In study of COVID-19 cases in Germany, an apparent discrepancy between cases and deaths was attributed to shifting age distribution, testing capacity, or true effectiveness of government-issued directives. While these country-level studies identify the three “artifacts” (H1-H3) as limitations of interpreting the CFR, none explicitly account for them.

To account for changes in testing capacity, we examine hospitalization data. While (as far as we are aware) no nation-wide studies in the U.S. account for all three artifacts, some hospital systems have controlled for them by conducting age-stratified cohort studies. Among 5,121 hospitalized COVID-19 patients in a single New York health system, Horwitz et. al. demonstrated that after adjusting for age, sex, ethnicity, and other clinical factors, HFR between March 1st and June 20th decreased but not as much as observed before adjusting for these factors. In another New York hospital system, Mehta et. al. demonstrated that cancer and older age were associated with increased risk of case fatality, finding no significant associations between race and mortality or gender and mortality. In a study conducted among 21,082 COVID-19 patients admitted to 108 English critical care units between March 1st and June 27th, mortality risk in mid-April and May was found markedly lower than earlier in the pandemic even after adjusting for age, sex, ethnicity, comorbidities, and geographic region. While these studies provide thorough estimates of mortality for their respective regions during their specific time periods, we analyze data over a longer time range and larger scale (588,126 hospitalizations, 10.3 million cases) that purportedly captures all of Florida and most of the United States.

Our data does not contain information about viral mutations and viral loads (H5 and H6). However, a few studies have begun to investigate their impact. The B.1.1.7 and B.1.351 variants of COVID-19 were first reported in the U.S. at the end of December 2020 and January 2021, respectively. While it is unclear how these variants will ultimately impact HFRs, recent studies indicate that vaccines may be more effective against the B.1.1.7 variant than the B.1.351 variant. Regarding social distancing precautions reducing viral load, in a study across seven countries, declining CFR was found to be correlated with strict lockdown policies and widespread PCR testing. At a hospital system in northern Italy, Piubelli et. al. found that among patients diagnosed with COVID-19 in their emergency room, the proportion of patients requiring intensive care decreased over time, also having lower viral load. Our analysis does not attempt to separate out the effects of viral mutations (H5) and changing viral loads (H6), but we note that these are factors that can affect the true infection fatality rate, and therefore can be reflected in our estimates as well.

Methodologically, prior studies on the COVID-19 fatality have employed logistic regression, Cox proportional hazards, and propensity matching to adjust for age and other comorbidities. While logistic regression and propensity matching can quantify risk of death averaged over their study time period, we are interested in the evolution of fatality risk over time. While the standard Cox proportional hazards does model risk over time, it assumes a simple linear form between the covariates and the log hazard. Without making this assumption, we leverage techniques in time series literature to reduce noise in the raw signal, and compute uncertainty around the estimates at any given time. Assuming smoothness in the true underlying trend, the moving average can obtain a better estimate of the trend than the raw signal. To quantify uncertainty in time series, the moving block bootstrap technique was developed in lieu of standard bootstrapping targeting independent and identically distributed observations. In this technique, the time series is chunked into blocks to reduce dependence among them. This reduced dependence is hard to ensure, however, thereby suggesting a strategy between model-based and block resampling. To reduce much of the dependence between original observations, one can “pre-whiten” the data by fitting a model to the data, and computing the residuals. Instead of block resampling the original dependent series, the residuals can be resampled, and added back to the model estimates. This intermediate solution, termed post-blackening, has been shown to work more consistently in practice.

Materials and Methods

Data Description We center our analysis on (1) state-level COVID-19 Case Line Data made available by the Florida Department of Health (FDOH) and (2) national-level COVID-19 Case Surveillance Data made available by the United States Centers for Disease Control and Prevention (CDC). Both datasets are line-level, including date of detection, demographics (including age and gender), and indicators of eventual outcomes for each case (Table 1). In our defined
cohorts, all COVID-19 cases are confirmed with a positive PCR lab result. Each FDOH case is marked with the date it was confirmed, and each CDC case is marked with the date it was reported. Overall, there are 1,004,818 confirmed cases in the FDOH data, and 10,332,725 in the CDC data.

Signal Smoothing To smooth out daily fluctuations in data reporting, for each date, we compute the 7-day lagged averages for COVID-19 cases, hospitalizations, and deaths. From this point onward, any discussion of these quantities or fatality rates based on them, unless otherwise stated, refers to the smoothed signal. Thus, in both FDOH and CDC data, we collect data extending back to March 26th in order to analyze the April 1st to December 1st time range.

Separating Artifacts from True Improvements. We argue that three main phenomena fuel an “artificial” decrease in CFR: increased testing capacity (H1), shifting age distributions (H2), and delays between detection and fatality (H3).

Since testing (H1) is not included in the FDOH or CDC data, we pull in data from The COVID Tracking Project to quantify increased testing capacity in Florida and nationally. We plot 7-day lagged averages of tests administered and the positive test rates. To avoid artifacts from increased testing, we examine changes in HFRs rather than CFRs.

To establish and account for shifting age distributions (H2), we examine cases, hospitalizations, and deaths stratified by age groups: 0-9, 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, and 80+. Naturally, age stratification reduces the amount of data for each estimate, so we omit HFR estimates that are based on fewer than two deaths.

Finally, to account for delays between detection and fatality (H3), we take advantage of the line-level nature of the data in order to perform a cohort-based analysis. For each date, we extract the cohort of individuals confirmed positive on that date, as well as whether those individuals eventually died or were hospitalized. By contrast, publicly reported case fatality rates are typically not cohort-based—the patients whose deaths are reported in the numerator are not in general the same patients whose confirmed infections show up in the denominator. Because case confirmation tends to precede reported deaths, these signals tend to be misaligned and are subject to fluctuation, even if the actual case fatality rate were fixed (so long as incidence does change). Line-level data enables us to circumvent this problem.

Taking the above three adjustments into account, our primary quantity of interest for treatment improvements is the age-stratified HFR. For the rest of the paper, we define CFR and HFR at day \( t \) as follows:

\[
\text{CFR}_t = \frac{\text{cases at day } t \text{ that eventually die}}{\text{cases at day } t}, \quad \text{HFR}_t = \frac{\text{cases at day } t \text{ that eventually get hospitalized and die}}{\text{cases at day } t \text{ that eventually get hospitalized}}
\]

Quantifying True Improvements Thus far, news and academic sources have highlighted three main “true improvements”: improvements in treatment (H4), disease mutations (H5), and reduced viral loads due to social distancing (H6). We seek to quantify treatment improvements (H4) by computing the drop in HFR.
Estimation using Block-Bootstrap and Cubic Splines. We use a cubic spline to fit the trend of the 7-day lagged average HFRs, and use a moving block-bootstrapping technique with post-blackening to estimate uncertainty around this trend. As described in the related work, 7-day lagged averages provide better estimate of trend by assuming smoothness rather than a specific functional form. Block-bootstrapping with post-blackening enables us to estimate uncertainty around this trend, with a weaker assumption than the standard i.i.d. assumption. We use a 7-day block size, based on the length of the time series and our observations that reporting follows a weekly cadence. After block-bootstrap resampling the residuals, the residuals are added back to the cubic spline, creating the replicates needed for estimating HFR with uncertainty. A visual walkthrough of this procedure is in the “HFR estimation” section of our web tool.

Visualization Tool. We publish an interactive web tool, available at acmilab.org/unpack_cfr, for dynamically applying our analyses to any demographic or date range of interest. On both FDOH data and CDC data, it displays plots for aggregate and age-stratified cases, hospitalizations, and deaths over time (Figure 3); plots for age distributions of cases, hospitalizations, and deaths over time (Figure 5); and estimates of age-stratified HFR as well as the change between two user-provided dates (Table 2 and 3). The user can select gender, race/ethnicity, and state to form cohorts of interest. For HFR estimates, the user can use a date selector to obtain new estimates for their date range of interest.

Results

Cases, Hospitalizations, and Deaths
In both the FDOH and the CDC data, one can discern three waves of COVID-19 cases. The first wave peaks in mid-April, the second wave peaks in mid-July, and cases leading up to December indicate a third wave (Figure 3).

Increased Testing. Between April 1st and December 1st, testing increases significantly, by approximately 964% in Florida and 1080% nationally (Figure 2). Florida observes a spike in testing near the second peak, whereas national testing rises more smoothly. We note that although peaks in Florida testing (Figure 2, left) coincide with spikes in Florida cases (Figure 3a, left), these spikes in cases cannot be entirely attributed to increased testing because there are also rises in positive test rates in April, July, and December (Figure 2, right).

Cases. Across all age strata, as measured by cases, Florida’s second wave is the most severe (Figure 3a, left) out of the three waves. In aggregate, it has approximately 1153% more cases than in the first peak and 46% more cases than in the ongoing third wave (Figure 3a, left). In contrast, nationally the third wave has substantially more cases than the first two peaks—392% more than the first peak and 150% more than the second peak (Figure 3b, left). Also, note that the relative jump in cases between the first two peaks is 96%, much less than the 1153% jump seen in Florida. This could be due to a combination of the spike in Florida’s testing in the second peak, as well as variation in the trajectories of different states (e.g. the populous state of New York was particularly hard-hit in the first wave).

Hospitalizations and Deaths. Overall, hospitalizations and deaths corroborate the story told by positive test rate (Figure 3a, center and right). In Florida, hospitalizations and deaths indicate a more severe second peak than first peak, though the contrast in peak size is not as dramatic as in the plot of cases. By contrast, in the national data, the second peak is smaller than the first. Much of the discrepancies of trends seen in cases versus in hospitalizations and deaths are likely attributable to increases in testing (Figure 2). Towards the third wave in December, Florida hospitalizations and deaths are at similar levels to that of the first wave. Nationally, the third wave appears to be worse than the second.

![Figure 2: COVID-19 positive test rates (right) and tests (left and middle) for Florida and the United States, calculated using 7-day trailing averages and pulled from the COVID Tracking Project. Positive test rate is calculated by dividing new positives by total new tests on each day. Data outside the April 1st to December 1st time range considered in this study is grayed out.](image-url)
Figure 3: Age-stratified cases, (eventual) deaths, and (eventual) hospitalizations in Florida and in the U.S., by the date of first positive test result (Florida) and date of report to the CDC (U.S.). Note that the x axis is not the date of death or date of hospitalization.

Figure 4: Aggregate case fatality rate (left) and hospitalization fatality rate (right) in Florida and in the U.S., by the date of first positive test result (Florida) or date of report to the CDC (U.S.).

Aggregate Fatality Rates. As expected, aggregate CFRs are lower than aggregate HFRs both in Florida and nationally (Figure 4). Unlike the national CFR which primarily decreases across the time range, the Florida HFR oscillates and reaches similar levels in both April and July. A deeper analysis of age-stratified fatality follows in later sections.

Confounding Due to Demographic Shift

Age. Between the first two peaks, the age distribution of cases shifts substantially, with the median age in Florida falling from 51 to 40, and the median national age group, from 50-59 to 30-39. After the second peak, the age distributions continue to fluctuate. In September, younger cases increase, possibly related to the start of the school year (Figure 5, left). By December 1st, the Florida median age remains at 40 but the national median age group rises to 40-49. Older individuals comprise a disproportionate share of hospitalization and death counts (Figure 5, middle and right).

Gender. The gender ratios in each age group’s cases, hospitalizations, and deaths appear relatively flat over time. Thus, in this paper we choose not to stratify by gender due to the reasonably small shift in the gender distribution over time, and practically to have more support in each group. However, we do provide this option in our web tool.

Age-stratified HFR

Consistent with prior literature, we find that as the age group increases so does the corresponding HFR (Tables 2 and 3). Measuring treatment improvements by HFR drop (computed as $\frac{HFR_{new} - HFR_{old}}{HFR_{old}}$), we observe larger treatment
Figure 5: Age distributions among Florida and national cases, (eventual) hospitalizations, and (eventual) deaths, by the date of first positive test result (Florida) and date of report to the CDC (U.S.), respectively.

Table 2: Estimates of HFR and drop in HFR on peak dates. Median and 95% confidence intervals are computed using block bootstrapping. Results with inadequate support are omitted.

<table>
<thead>
<tr>
<th>Age group</th>
<th>Florida</th>
<th>National</th>
</tr>
</thead>
<tbody>
<tr>
<td>aggregate</td>
<td>0.23 (0.21, 0.26)</td>
<td>-0.023 (-0.14, 0.13)</td>
</tr>
<tr>
<td>20-29</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>30-39</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>40-49</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>50-59</td>
<td>0.092 (0.078, 0.11)</td>
<td>0.12 (-0.085, 0.38)</td>
</tr>
<tr>
<td>60-69</td>
<td>0.18 (0.15, 0.21)</td>
<td>0.13 (-0.045, 0.38)</td>
</tr>
<tr>
<td>70-79</td>
<td>0.31 (0.28, 0.34)</td>
<td>0.33 (0.31, 0.34)</td>
</tr>
<tr>
<td>80+</td>
<td>0.46 (0.43, 0.49)</td>
<td>0.48 (0.46, 0.49)</td>
</tr>
</tbody>
</table>

improvements between April and December to be correlated with younger age (Table 3). Note, however, that the younger groups have small HFRs to begin with, so the opposite trend might appear when considering absolute rather than relative improvements. Additionally, the confidence intervals for HFR drops are wider for younger age groups.

Between the first two peaks (Table 2), the national age-stratified HFR estimates from block bootstrapping decrease by as little as 27% in the 80+ age group, and as much as 37% in the 30-39 age group. On the other hand, in Florida the age-stratified HFR actually increases in each age group by as little as 2.9% in the 80+ age group, and as much as 13% in the 60-69 age group. Note that the HFR changes between peak dates in Florida are an example of Simpson’s paradox, where in each age group the HFR increase, but the aggregate HFR actually decreases by 2.3%.

Compared to peak-to-peak changes, across the entire time range (Table 3) we observe a more dramatic decrease in HFR. In Florida, the HFR drops by as little as 17% in the 80+ age range, and as much as 42% in the 60-69 age range. Nationally, the HFR drops by as little as 55% in the 80+ age groups, and as much as 73% in the 20-29 age group.

While this paper presents estimates at the two peaks and the endpoints of the study time range, we can easily read off similar estimates with uncertainty for all dates between April 1st and December 1st. This type of interactive functionality is available in our web tool. In our web tool, when stratifying by gender in addition to age, the conclusions surrounding drops in HFR are similar to those when just stratifying by age.
Table 3: Estimates of HFR and drop in HFR between April 1st and December 1st. Median and 95% confidence intervals are computed using block bootstrapping. Results with inadequate support are omitted.

<table>
<thead>
<tr>
<th>Age group</th>
<th>Florida</th>
<th>National</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2020-04-01</td>
<td>2020-12-01</td>
</tr>
<tr>
<td>aggregate</td>
<td>0.23 (0.2, 0.27)</td>
<td>0.16 (0.12, 0.19)</td>
</tr>
<tr>
<td>20-29</td>
<td>0.18 (0.16, 0.2)</td>
<td>-</td>
</tr>
<tr>
<td>30-39</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>40-49</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>50-59</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>60-69</td>
<td>0.18 (0.14, 0.22)</td>
<td>0.1 (0.06, 0.14)</td>
</tr>
<tr>
<td>70-79</td>
<td>0.31 (0.26, 0.35)</td>
<td>0.19 (0.14, 0.23)</td>
</tr>
<tr>
<td>80+</td>
<td>0.44 (0.39, 0.49)</td>
<td>0.37 (0.32, 0.41)</td>
</tr>
</tbody>
</table>

Discussion

We unpack the drop in CFR to quantify improvements reasonably attributable to advances in treatment, accounting for shifting age distributions (H1) by age-stratifying, increased testing capacity (H2) by focusing on the hospitalized, and the detection-to-fatality delay (H3) by conducting a cohort-based analysis. We find that increased testing does not explain away the three waves due to corresponding peaks in hospitalizations, deaths, and positive test rates. We visualize the shifting age distributions, and quantify the decrease in age-stratified HFRs between the first two peaks and across the entire study time range. Combining all these analyses, we arrive at the following narrative:

At the beginning of April, testing was relatively sparse (Figure 2). Cases, hospitalizations, and deaths were rising, and reached peak levels circa April 15th (Figure 3). Roughly one in every ten tests came back positive in Florida, and one in every five tests, nationally. In Florida, the aggregate HFR was approximately 23%, with age-stratified HFRs ranging between 9.2% for the 50-59 age group to 46% for the 80+ age group (Table 2). Nationally, the aggregate HFR was approximately 29%, with the age-stratified HFRs ranging between 21.1% for the 20-29 age group and 57% for the 80+ age group (Table 2). In each age group, the national HFR was higher than the Florida HFR, possibly due to overwhelmed hospital systems in states hit hard during the first wave. In fact, our web tool indicates that 34.5% of national CDC cases between April 1st and April 15th were recorded in New York alone.

Over the next three months, the proportion of younger individuals with COVID-19 grew steadily (Figure 5). Testing continued to rise nationally, and spiked in Florida as it approached a heavier second peak around July 15, with positive test rates also at an all-time high (Figure 2). Florida experienced record hospitalizations and deaths, and the age-stratified HFRs were at least as high as in the first wave (Table 2). While Bill Gates had publicly attributed “a factor-of-two improvement in hospital outcomes” to dexamethosone and remdesivir, this did not yet appear to be true in Florida. (Alternatively, treatment improvements might have been counterbalanced by strain on the hospital system.) On the other hand, cases in New York had diminished (shown in our web tool) and were starting to surge in other states, forming a smaller second peak nationally (as measured by hospitalizations and deaths). Between the first two peaks, the national HFR had dropped by 41% in aggregate, with age-stratified HFRs dropping as much as 37% in the 30-39 age group and as little as 27% in the 80+ age group. The different stories told here by Florida and the national aggregate data underscore the importance of state-level rather than national analysis.

Finally, come December 1st, a third wave is underway. Approximately 31% of all national cases since April were confirmed in the last month alone (Figure 3b). In terms of hospitalizations, deaths, and positive test rate, this third wave has already surpassed the second wave. For Florida, the third wave is already at least as severe as the first wave. Fortunately, age-stratified HFRs in both Florida and the national aggregate data appear to have dropped significantly since the start of the pandemic, likely indicating treatment improvements (though possibly confounded by H5 and H6). Since April 1st, the age-stratified HFR in Florida has decreased by as much as 42% in the 60-69 age group and as little as 17% in the 80+ age group. Nationally, the age-stratified HFR has decreased by as much as 73% in the 20-29 age group and as little as 55% in the 80+ group. Regarding the CFR, on July 27, former President Donald Trump stated in a press briefing that “Due to the medical advances we’ve already achieved and our increased knowledge in how to treat the virus, the mortality rate for patients over the age of 18 is 85 percent lower than it was in April.”

Note, however, that none of our estimates of treatment improvements are as large as the 85% touted by Trump. In summary, CFR can be misleading if age distribution shift, increased testing, and detection-to-death delay are unaccounted for.
Limitations

We aim to quantify treatment improvements (H4) in Florida and the U.S. by estimating changes in the age-stratified HFR, but H4 could also be influenced by disease mutation (H5) and changing viral loads (H6). To distinguish their effects in future work, we need additional data. Furthermore, while we listed the six hypotheses we found in literature review, possible alternative explanations may arise in the future as pandemic evolves.

We assume that treatment improvements will be reflected in the HFR because over our study’s time range, major treatment improvements (e.g., dexamethosone and remdesivir) targeted hospitalized patients. While the first U.S. vaccination was administered outside of our study time range, vaccines take effects before hospitalization, and so their treatment improvements may not be reflected in the HFR for future studies. While our method could still quantify post-hospitalization treatment improvements, we note that vaccination roll-out criteria (e.g. occupation, age) and other characteristics (e.g. socioeconomic background) could influence who gets hospitalized in the first place.

Other limitations arise from data quality issues. In both the FDOH and CDC data, hospitalization and death are highly missing (Table 1), possibly introducing bias in the HFR estimates if the data are not missing at random. States have different patterns when reporting data to the CDC (seen by filtering for any state in our web tool). The reported CDC cases appear to be incomplete. For instance, the CDC cases reported from Florida only account for 67% of the cases provided by the FDOH, they are reported sporadically even after smoothing, and no death is reported since October. Cross-referencing with the COVIDcast API, we find that in the CDC data reported from Texas, only 4.9% of the cases and 0.01% of the deaths are accounted for, and only 14 hospitalizations were recorded. Thus, in our national analysis, we are making the assumption that in aggregate the signal will outweigh the noise. Despite the data limitations, the CDC data appears to be the best available source of line-level cases needed for cohort-based analysis across the United States. We note that in the Florida FDOH data, on the other hand, we use the positive test confirmed date which is not missing at all in this data, making the Florida HFR estimates more reliable than those from the national data.

References

Transitions of Care: Completeness of the Interoperability Data Standard for Communication from Home Health Care to Primary Care

Edgar Chou, MD, MBA, FACP\textsuperscript{1}; Paulina S. Sockelow, DrPH, MS, MBA\textsuperscript{2}
\textsuperscript{1} Thomas Jefferson University, Philadelphia, PA; \textsuperscript{2} Drexel University, Philadelphia, PA

Abstract

Data sharing is necessary to address communication deficits along the transitions of care among community settings. Evidence-based practice supports home healthcare (HHC) patients to see their primary care team within the first two weeks of hospital discharge to reduce rehospitalization risk. A small subset of patient data collected at HHC admission is mandated to be transmitted to primary care, predominantly by fax. Using qualitative analysis, we assessed completeness of the United States Core Data for Interoperability (USCDI) interoperability standard, as compared to the patient data collected by the primary care team (topics) and HHC (classes) during the initial visit; and offer interoperability recommendations. Findings indicate the USCDI does not cover 74\% of the 19 faxed HHC classes that mapped to the primary care topics, and 95\% of the 38 not-faxed HHC classes. We offer USCDI recommendations to address these interoperability gaps.

Keywords: communication, home health care nursing, home health nursing, primary health care, continuity of patient care/standards, transition of care, nursing informatics, documentation

Introduction

Data sharing across transitions of care is necessary to address fragmented care delivery along the health care continuum. Much of the patient care is provided in community settings where, due to information silos and lack of interoperability, data communication deficits exist along the transition of care among community settings. These deficits are becoming more apparent as health care organizations move towards value-based care where hospitalizations and readmissions are key quality and cost drivers. To support data sharing, the United States Core Data for Interoperability (USCDI) replaces the Common Clinical Data Set (CCDS)\textsuperscript{1} as a data standard in transitions of care.\textsuperscript{2} USCDI support for data communication among community settings warrants investigation.

Home healthcare services (HHC)\textsuperscript{3} and a primary care team visit are transition-in-care components for over one million Medicare patients annually discharged from hospital to home. Evidence-based practice supports HHC patients seeing their primary care team within the first two weeks of hospital discharge\textsuperscript{4,5} which has been shown to reduce rehospitalization risk in patients with heart failure\textsuperscript{6} and sepsis.\textsuperscript{7} Accordingly, the Centers for Medicare and Medicaid (CMS) reimburses providers for a patient office visit within 14 days of hospital discharge, the Transition of Care (TOC) visit.\textsuperscript{8} Patients with medical and/or psychosocial problems requiring at least moderate medical decision-making receive services\textsuperscript{9} which results in lower mortality and cost of care.\textsuperscript{8} Evidence and CMS regulations highlight TOC visit importance. Left unsaid are the specifics of what information is needed for the TOC visit – for medical decision-making and prospective analysis e.g., re-hospitalization risk.

A rich set of information relevant to TOC clinical decision-making is collected during the HHC visit as structured data yet little of this data is transmitted to the primary care team. During the first (admission) visit HHC clinicians conduct medication reconciliation, assessment, and documentation of patient cognitive and functional capabilities and patient safety in the CMS Outcome and Assessment Information Set (OASIS\textsuperscript{10}). To finalize the admission, a minimal OASIS data subset is faxed to the primary care physician to request clinician order sign-off for HHC reimbursement. The faxed form (the ‘485’) is difficult to interpret due to the paucity of structured data, preponderance of narrative text, and asynchronicity of receipt in relation to the TOC patient visit. The brevity of the typically 20-minute TOC visit limits the information physicians can gather as they address acute and chronic illnesses. Yet important information is in the OASIS. For example, a recent literature review identified OASIS data that predicts re-hospitalization risk.\textsuperscript{11} Another example is HHC identifies a patient living with cognitive deficits which affect medication ingestion. This information could inform the primary care team of the need for additional resources to ensure the patient’s safe pill consumption. This communication occurs verbally, if at all: No formatted or structured documentation of this information, paper or electronic, reaches the physician office.

Lack of HHC and primary care EHR interoperability promulgates information siloes. The result is potential missed clinical opportunities (not having information in the right place at the right time) and lack of data for research such as re-hospitalization risk predictive analytics. Data flow deficits between HHC and primary care can impact patient
outcomes, and hospital and practice CMS reimbursement. The USCDI may address the needed interoperability from HHC to primary care. We undertook an investigation of USCDI completeness as compared to HHC patient information (a) currently communicated to primary care through the 485 form and mapped to the TOC; and (b) which could be communicated from OASIS for the TOC visit. From this analysis we seek to provide HHC-to-primary care interoperability recommendations.

Methods

The authors conducted deductive qualitative analysis to meet the objectives in two steps. First, they assessed the extent to which HHC patient data currently communicated on the 485 provides the information needed for the TOC, and the completeness of the USCDI relative to the 485 data that mapped to the TOC. Second, they mapped OASIS data not communicated on the 485 to the TOC, and assessed completeness of the USCDI relative to the OASIS data that mapped to the TOC. The current version of OASIS, Version D, was used in the analysis. The authors’ perspective was that of a primary care physician knowledgeable about HHC OASIS data who was considering the HHC information the primary care team might need. The perspective was not that of a physician more narrowly interpreting the TOC topics as it is currently implemented. An example of the former is medication self-administration issues, and example of the latter is medication list. The chosen perspective was selected to provide insight into the potential value of HHC data to the primary care team.

Completeness of USCDI coverage of 485 data mapped to the TOC

In the absence of a CMS TOC document, we used a primary care medical society TOC. It contained information descriptions (referred herein as topics) and lacked structured data fields. This analysis focused on TOC activities or decisions which require data (e.g., Obtain and Review Discharge Information). TOC topics could be interpreted as having some overlap, for example, implying review of the same data (e.g., Obtain and Review Discharge Information, and Review Need for Follow-up on Pending Testing or Treatment both include review of clinical status). The TOC contained 9 clinical information topics distributed among 3 domains as follows. The four clinical status topics were:

- Obtain and Review Discharge Information
- Review Need for Follow-up on Pending Testing or Treatment
- Interact with Other Clinicians who will Assume or Resume Care of the Patient’s System-specific Conditions
- Establish or Re-establish Referrals for Specialized Care

The two functional status topics were:

- Educate the Patient and/or Caregiver to Support Self Management, and Activities of Daily Living (ADL)
- Provide Assessment and Support for Treatment Adherence and Medication Management

The three service needs topics were:

- Identify Available Community and Health Resources
- Facilitate Access to Services Needed by the Patient and/or Caregivers
- Assist in Scheduling Follow-up with Other Health Services

The topic, Communicate with Agencies and Community Services Used by the Beneficiary, was excluded from the analysis as it is not currently interpreted as referring to 485 information due to the asynchronicity.

The 485 communicates 19 specified structured and unstructured data (classes). Information sources are 13 OASIS classes for the structured data, and unstructured, non-OASIS EHR data for 10 classes. Some 485 classes have multiple OASIS classes as data sources. The classes are in 3 domains: clinical status, functional status, and safety. The thirteen 485 clinical status classes include 9 OASIS-sourced classes (i.e., Primary Diagnosis/Other Diagnoses [3 questions], Cognitive, Behavioral and Psychiatric Symptoms [4 questions], Living Arrangements [1 question], Hospitalization Risk [1 question]) and 5 EHR-sourced classes (i.e., Orders for Discipline and Treatments, Prognosis, Medications List, Allergies, Nutritional Requirements). The 5 functional status classes include 4 OASIS-sourced classes (i.e., Goals/Rehabilitation Potential/Discharge Plans) and 4 EHR-sourced classes (i.e., Mental Status, Functional Limitations, Activities Permitted, Durable Medical Equipment [DME] and Supplies). Note that Goals are not specified as to whether patient or clinician goal. The single safety class was from EHR data (i.e., Safety Measures).

The USCDI contains standards at incremental levels of adoption: Version 1, Draft Version 2, Level 2, Level 1 and Comments. Version 2 is the focus of this analysis as it is the most current proposed standard and incorporates Version 1 which has been adopted. Levels suggest data elements that are in discussion to be incorporated in subsequent
versions. Comments contain data elements that have adoption hurdles, including insufficiently defined use cases and implementation or development burdens.\textsuperscript{13} Excluded from the analysis, but of note, Comments includes functional class (e.g., Mental, Mobility, Self Care, and Instrumental ADL [IADL]). The USCDI contains data classes and elements. Elements which were qualifiers for classes (e.g., dosage as a qualifier for medication) were excluded from the analysis. Referring to ONC documentation,\textsuperscript{13} the authors reviewed classes to identify those relevant to HHC and TOC patient information. Excluded USCDI classes were: Provenance, Implantable Device Identifier, and Patient Demographics. The retained USCDI classes were 14 structured classes, all in the clinical domain, including Medications List, Problems, Procedures, (Patient) Goals, and Care Team Member.

The authors independently extracted topics (TOC) and classes present on the 485 form and used an Excel spreadsheet (Microsoft Corporation, Redmond, WA) for organization. The authors compared their extractions and resolved differences. They decided to avoid redundant mapping due to overlapping TOC topics by mapping 485 data to the more closely mapped TOC topic. Drawing on the physician author’s clinical expertise and working collaboratively, the authors categorized the extracted topics and classes from each source into domains present on either source (i.e., clinical status, functional status, home safety) and classes within each domain (e.g., for clinical status: Medications List, Problems). For example, the domain clinical status included the following classes within the topic of medications from the designated sources: TOC- provide assessment and support for treatment adherence and medication management; 485- medications; USCDI- medications. The class/element hierarchy is consistent with that of the USCDI. The authors also characterized each class from each source as structured or unstructured.

The authors, in partnership, identified topics and classes that mapped or did not map between the TOC and the 485. The working description of mapping was if a class was included fully or partially in a topic. An example of a partial mapping is that discharge information includes many classes. Therefore, instead of a single class, many classes (e.g., medication issues, fall risk, ADLs) provide information to the TOC topic, Obtain and Review Discharge Information. Each 485 class was reviewed for mapping to one or more TOC topic which shared the underlying concept. For example, the 485 class ‘Medications List’ mapped to the TOC topic Obtain and Review Discharge Information which infers inclusion of problems, medications, and allergies. The analysis produced a list of topics and classes, each characterized as mapped or unmapped, and tagged with data source (TOC, 485), domain, and structured/unstructured.

Working together, the authors mapped the listed topics and classes to the USCDI. The mapping indicated completeness of the data standard relative to the TOC information and the currently faxed HHC information. Incompleteness of the standard indicated areas for development of USCDI transitions of care interoperability recommendations.

Completeness of USCDI coverage of OASIS data mapped to the TOC

The OASIS contains 68 questions related to the admission. The authors eliminated 16 demographic and insurance questions, and 1 question lacking clinical relevance (inpatient discharge date). Retained were 51 patient health classes (questions) to assess the patient in domains including clinical status, functional status, home safety, and service needs.\textsuperscript{10} All OASIS data is structured. Most question responses are categorical, typically with 3-to-6-point scales. The 34 clinical status classes include Medication Issues, body systems (including emotional and behavioral), and Hospitalization Risk. The 13 functional status classes include ADL/IADL, and Self Management. The 2 home safety classes are Living Arrangements (regarding presence of other people) and Falls Risk. The 2 service needs classes focus on care management related to supervision needed for safety.

Together the authors compared the 485 and OASIS classes, retaining classes unique to OASIS (non-unique classes were mapped above). The examination of retained classes followed the data collection and analysis process described above to identify classes that mapped (fully or partially) and did not map between the TOC and OASIS. For example, the OASIS class Medication Issues did not appear on the 485 and mapped to the TOC topic Obtain and Review Discharge Information. Mapping to the USCDI indicated completeness of the USCDI relative to HHC OASIS data available and not communicated to primary care, suggesting areas for USCDI recommendation development.

Results

The comparison of TOC topics, 485 classes, and USCDI classes is shown in Table 1. Results of the comparison are described below.

Completeness of USCDI coverage of 485 data mapped to the TOC

Comparison of TOC topics and 485 classes indicated all 19 of the 485 classes mapped to 6 TOC topics, and 3 TOC topics were unmapped. Nine 485 classes mapped to multiple TOC topics, and 4 TOC topics had multiple 485 classes which mapped to them. The nine 485 classes that mapped to multiple TOC topics were: Goals/Rehabilitation

297
Potential/Discharge Plans, Medications List, Hospitalization Risk, Living Arrangements, Orders for Discipline and Treatments, Functional Limitations, Activities Permitted, DME and Supplies, and Safety Measures. The 4 TOC topics with multiple 485 classes were: Obtain and Review Discharge Information (all 485 classes); Establish or Reestablish Referrals for Specialized Care (3 classes); Educate the Patient and/or Caregiver to Support Self Management and ADL (5 classes); and Identify Available Community and Health Resources (3 classes). Two TOC topics each had a single 485 class which mapped to it: Review Need for Follow-up on Pending Testing or Treatment, and Provide Assessment and Support for Treatment Adherence and Medication Management. The 3 unmapped TOC topics were: Interact with Other Clinicians who will Assume or Resume Care of the Patient’s System-specific Conditions, Assist in Scheduling Follow-up with Other Health Services, and Facilitate Access to Services Needed by the Patient and/or Caregivers.

Contrasting TOC topics and USCDI classes indicated all 14 USCDI classes mapped to 5 TOC topics, and 4 TOC topics were unmapped. Eight USCDI classes mapped to multiple TOC topics, and 3 TOC topics had multiple USCDI classes which mapped to them. The 8 USCDI classes that mapped to multiple TOC topics were: Clinical Notes, Medications List, Assessment and Plan of Treatment, Laboratory, Procedures, Diagnostic Imaging, Health Concerns, and (Patient) Goals. The 3 TOC topics with multiple USCDI classes were: Obtain and Review Discharge Information (all USCDI classes); Review Need for Follow-up on Pending Testing or Treatment (6 classes); and Establish or Reestablish Referrals for Specialized Care (6 classes). The two TOC topics with a single mapped USCDI class were: Provide Assessment and Support for Treatment Adherence and Medication Management; and Educate the Patient and/or Caregiver to Support Self-Management and ADL. The 4 unmapped TOC topics were: Interact with Other Clinicians who will Assume or Resume Care of the Patient’s System-specific Conditions, and all the service needs topics: Identify Available Community and Health Resources, Facilitate Access to Services Needed by the Patient and/or Caregivers, and Assist in Scheduling Follow-up with Other Health Services.

Looking across the TOC, 485, and USCDI mappings, the single TOC topic to which the 485 classes mapped, but remained unmapped by USCDI classes, was Identify Available Community and Health Resources. While all the 485 classes mapped to the TOC, only 5 of 19 classes (26%) mapped to the USCDI. Four of these five 485 classes directly overlapped with the USCDI classes of Goals, Problems, Medications List, and Allergies. The 485 class Goals/Rehabilitation Potential/Discharge Plans provided additional data beyond the USCDI class Goals. Fourteen 485 classes did not map to the USCDI: 13 were structured and 1 was unstructured (Orders for Discipline and Treatments). The structured classes spanned the clinical (9), functional (4), and safety (1) domains. The unstructured class was in the clinical domain.

Completeness of USCDI coverage of OASIS data mapped to the TOC

Examination of the OASIS and the 485 indicated that of the 51 OASIS classes, 13 (25%) mapped to the 485. The mapped classes were distributed in the following domains: clinical (8), functional (4) and safety (1).

Comparison of TOC topics and the remaining 38 OASIS classes indicated all remaining OASIS classes mapped to 6 TOC topics, and 3 TOC topics were unmapped. Nineteen OASIS classes (50% of remaining classes) mapped to multiple TOC topics, and four TOC topics had multiple OASIS classes which mapped to them. Herein we refer to related OASIS codes classes groupings, e.g., the 5 medication related classes are referred to as Medication Issues. The number of OASIS classes related to a grouping is noted in parentheses. The 8 groupings of the 19 OASIS classes that mapped to multiple TOC topics were: History and Physical (1 OASIS class); Home Therapies (1 class); ADL/IADLs (9 classes); Falls Risk (1 class); Medication Issues (5 classes); Care Management (1 class); and Therapy Need (1 class). The 5 TOC topics with multiple OASIS classes were: Obtain and Review Discharge Information (all OASIS classes); Establish or Re-establish Referrals for Specialized Care (4 classes); Educate the Patient and/or Caregiver to Support Self Management and ADL (9 classes); Provide Assessment and Support for Treatment Adherence and Medication Management (6 classes); and Identify Available Community and Health Resources (4 classes). One TOC topic had a single OASIS class which mapped to it: Review Need for Follow-up on Pending Testing or Treatment. The 3 TOC topics that were unmapped in the 485 analysis remained unmapped in the OASIS analysis.

Looking at the possible data flow of OASIS classes to the TOC through the USCDI indicated two findings. Of the 38 OASIS classes not on the 485, two (5%, i.e., History & Physical, Height & Weight) mapped to USCDI classes (i.e., Clinical Notes, Vital Signs). For example, the History & Physical class in OASIS is one component of the Clinical Notes class in the USCDI. In total, 42 OASIS classes (83% of all OASIS classes) did not map to USCDI. They spanned the clinical (29; 57% of all classes), functional (9; 18%), service needs (2; 4%) and safety (2; 4%) domains. All classes are structured.
Table 1. USCDI Coverage of the TOC and Home Health Care Data

<table>
<thead>
<tr>
<th>TOC Topic(s)</th>
<th>485 Class</th>
<th>OASIS Group (Number of Classes)</th>
<th>USCDI Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obtain and Review Discharge Information; Establish or Re-establish Referrals for Specialized Care; Educate the Patient and/or Caregiver to Support Self-Management and Activities of Daily Living</td>
<td>Goals/Rehabilitation Potential/Discharge Plans</td>
<td>Functional Abilities and Goals (4)</td>
<td>(Patient) Goals</td>
</tr>
<tr>
<td>Obtain and Review Discharge Information</td>
<td>Primary Diagnosis/ Other Diagnoses (2)</td>
<td>Primary Diagnosis (3)</td>
<td>Problems</td>
</tr>
<tr>
<td>Obtain and Review Discharge Information; Provide Assessment and Support for Treatment Adherence and Medication Management</td>
<td>Medications List</td>
<td>No Match</td>
<td>Medications List</td>
</tr>
<tr>
<td>Obtain and Review Discharge Information</td>
<td>Allergies</td>
<td>No Match</td>
<td>Allergies &amp; Intolerances</td>
</tr>
<tr>
<td>Obtain and Review Discharge Information; Review Need for Follow-up on Pending Testing or Treatment; Establish or Re-establish Referrals for Specialized Care</td>
<td>No Match</td>
<td>History and Physical (1)</td>
<td>Clinical Notes</td>
</tr>
<tr>
<td>Obtain and Review Discharge Information</td>
<td>No Match</td>
<td>No Match</td>
<td>Immunizations</td>
</tr>
<tr>
<td>Obtain and Review Discharge Information</td>
<td>No Match</td>
<td>Height and Weight (1)</td>
<td>Vital Signs</td>
</tr>
<tr>
<td>Obtain and Review Discharge Information; Review Need for Follow-up on Pending Testing or Treatment; Establish or Re-establish Referrals for Specialized Care</td>
<td>No Match</td>
<td>No Match</td>
<td>Health Concerns</td>
</tr>
<tr>
<td>Obtain and Review Discharge Information; Review Need for Follow-up on Pending Testing or Treatment; Establish or Re-establish Referrals for Specialized Care</td>
<td>No Match</td>
<td>No Match</td>
<td>Assessment and Plan of Treatment</td>
</tr>
<tr>
<td>TOC Topic</td>
<td>485 Class</td>
<td>OASIS Group (Number of Classes)</td>
<td>USCDI Class</td>
</tr>
<tr>
<td>--------------------------------------------------------------------------</td>
<td>----------------------</td>
<td>--------------------------------</td>
<td>-------------------</td>
</tr>
<tr>
<td>Obtain and Review Discharge Information; Review Need for Follow-up on Pending Testing or Treatment; Establish or Re-establish Referrals for Specialized Care</td>
<td>No Match</td>
<td>No Match</td>
<td>Laboratory</td>
</tr>
<tr>
<td>Obtain and Review Discharge Information; Review Need for Follow-up on Pending Testing or Treatment; Establish or Re-establish Referrals for Specialized Care</td>
<td>No Match</td>
<td>No Match</td>
<td>Procedures</td>
</tr>
<tr>
<td>Obtain and Review Discharge Information; Review Need for Follow-up on Pending Testing or Treatment; Establish or Re-establish Referrals for Specialized Care</td>
<td>No Match</td>
<td>No Match</td>
<td>Diagnostic Imaging</td>
</tr>
<tr>
<td>Obtain and Review Discharge Information</td>
<td>No Match</td>
<td>No Match</td>
<td>Care Team Members</td>
</tr>
<tr>
<td>Obtain and Review Discharge Information</td>
<td>No Match</td>
<td>No Match</td>
<td>Smoking Status</td>
</tr>
<tr>
<td>Obtain and Review Discharge Information; Establish or Re-establish Referrals for Specialized Care; Identify Available Community and Health Resources</td>
<td>Hospitalization risk</td>
<td>Hospitalization risk (1)</td>
<td>No Match</td>
</tr>
<tr>
<td>Obtain and Review Discharge Information; Establish or Re-establish Referrals for Specialized Care; Identify Available Community and Health Resources</td>
<td>Living Arrangements</td>
<td>Living Arrangements (1)</td>
<td>No Match</td>
</tr>
<tr>
<td>Obtain and Review Discharge Information</td>
<td>Neuro/Emotional/Behavioral Status</td>
<td>Neuro/Emotional/Behavioral Status (1)</td>
<td>No Match</td>
</tr>
<tr>
<td>Obtain and Review Discharge Information</td>
<td>When is patient confused</td>
<td>When is patient confused (1)</td>
<td>No Match</td>
</tr>
<tr>
<td>Obtain and Review Discharge Information</td>
<td>When is patient anxious</td>
<td>When is patient anxious (1)</td>
<td>No Match</td>
</tr>
<tr>
<td>TOC Topic</td>
<td>485 Class</td>
<td>OASIS Group (Number of Classes)</td>
<td>USCDI Class</td>
</tr>
<tr>
<td>---------------------------------------------------------------------------</td>
<td>---------------------------------------------------------------------------</td>
<td>------------------------------------------------------------------------------------------------</td>
<td>--------------</td>
</tr>
<tr>
<td>Obtain and Review Discharge Information</td>
<td>Cognitive, Behavioral, and Psychiatric Symptoms</td>
<td>Cognitive, Behavioral, and Psychiatric Symptoms (1)</td>
<td>No Match</td>
</tr>
<tr>
<td>Obtain and Review Discharge Information; Review Need for Follow-up on Pending Testing or Treatment</td>
<td>Orders for Discipline and Treatments</td>
<td>No Match</td>
<td>No Match</td>
</tr>
<tr>
<td>Obtain and Review Discharge Information</td>
<td>Prognosis</td>
<td>No Match</td>
<td>No Match</td>
</tr>
<tr>
<td>Obtain and Review Discharge Information</td>
<td>Mental Statuses</td>
<td>No Match</td>
<td>No Match</td>
</tr>
<tr>
<td>Obtain and Review Discharge Information</td>
<td>Nutritional Requirements</td>
<td>No Match</td>
<td>No Match</td>
</tr>
<tr>
<td>Obtain and Review Discharge Information; Educate the Patient and/or Caregiver to Support Self-Management and Activities of Daily Living</td>
<td>Functional Limitations</td>
<td>No Match</td>
<td>No Match</td>
</tr>
<tr>
<td>Obtain and Review Discharge Information; Educate the Patient and/or Caregiver to Support Self-Management and Activities of Daily Living</td>
<td>Activities Permitted</td>
<td>No Match</td>
<td>No Match</td>
</tr>
<tr>
<td>Obtain and Review Discharge Information; Educate the Patient and/or Caregiver to Support Self-Management and Activities of Daily Living</td>
<td>Durable Medical Equipment (DME) and Supplies</td>
<td>No Match</td>
<td>No Match</td>
</tr>
<tr>
<td>Obtain and Review Discharge Information; Educate the Patient and/or Caregiver to Support Self-Management and Activities of Daily Living; Identify Available Community and Health Resources</td>
<td>Safety Measures</td>
<td>No Match</td>
<td>No Match</td>
</tr>
<tr>
<td>Obtain and Review Discharge Information; Provide Assessment and Support for Treatment Adherence and Medication Management</td>
<td>No Match</td>
<td>Home Therapies (1)</td>
<td>No Match</td>
</tr>
<tr>
<td>Obtain and Review Discharge Information</td>
<td>No Match</td>
<td>Sensory Status, Pain Impact on Activity (2)</td>
<td>No Match</td>
</tr>
<tr>
<td>TOC Topic</td>
<td>485 Class</td>
<td>OASIS Group (Number of Classes)</td>
<td>USCDI Class</td>
</tr>
<tr>
<td>--------------------------------------------------------------------------</td>
<td>--------------</td>
<td>--------------------------------------------------</td>
<td>--------------</td>
</tr>
<tr>
<td>Obtain and Review Discharge Information</td>
<td>No Match</td>
<td>Integumentary Status- ulcers (9)</td>
<td>No Match</td>
</tr>
<tr>
<td>Obtain and Review Discharge Information</td>
<td>No Match</td>
<td>Respiratory Status (1)</td>
<td>No Match</td>
</tr>
<tr>
<td>Obtain and Review Discharge Information</td>
<td>No Match</td>
<td>Elimination Status (4)</td>
<td>No Match</td>
</tr>
<tr>
<td>Obtain and Review Discharge Information</td>
<td>No Match</td>
<td>Depression Screening (1)</td>
<td>No Match</td>
</tr>
<tr>
<td>Obtain and Review Discharge Information</td>
<td>No Match</td>
<td>Disruptive Behavior (1)</td>
<td>No Match</td>
</tr>
<tr>
<td>Obtain and Review Discharge Information; Educate the Patient and/or Caregiver to Support Self Management and Activities of Daily Living</td>
<td>No Match</td>
<td>ADL/IADLs (9)</td>
<td>No Match</td>
</tr>
<tr>
<td>Obtain and Review Discharge Information; Establish or Re-establish Referrals for Specialized Care; Identify Available Community and Health Resources</td>
<td>No Match</td>
<td>Falls Risk (1)</td>
<td>No Match</td>
</tr>
<tr>
<td>Obtain and Review Discharge Information; Provide Assessment and Support for Treatment Adherence and Medication Management</td>
<td>No Match</td>
<td>Medication Issues (5)</td>
<td>No Match</td>
</tr>
<tr>
<td>Obtain and Review Discharge Information; Establish or Re-establish Referrals for Specialized Care; Identify Available Community and Health Resources</td>
<td>No Match</td>
<td>Care Management (1)</td>
<td>No Match</td>
</tr>
<tr>
<td>Obtain and Review Discharge Information; Establish or Re-establish Referrals for Specialized Care; Identify Available Community and Health Resources</td>
<td>No Match</td>
<td>Therapy Need (1)</td>
<td>No Match</td>
</tr>
</tbody>
</table>

**Discussion**

We examined transitions of care patient data needed in patient-centered primary care (TOC), available in home health care (OASIS) and communicated from HHC to primary care (485) to assess the completeness of the transitions of care.
data standard (USCDI). This analysis is operationally relevant: Conducting the TOC evaluation within the first two weeks of the HHC patient’s hospital discharge is a CMS requirement associated with reduced re-hospitalizations.\textsuperscript{8}

Findings indicate a deficit in the information needed for decision-making during the TOC visit. Data faxed from HHC did not map to three of the nine TOC topics. Furthermore, HHC data is unlikely to be available in the primary care TOC visit workflow. HHC information is not communicated electronically as structured data and can be sent days before and even after the TOC visit.

Our analysis also suggests a paucity of structured HHC patient data – likely to be informative to the primary care team – is currently communicated to primary care. Thirty-eight OASIS classes not currently faxed to primary care mapped to TOC topics. This result suggests the prescribed data flow from HHC to primary care (i.e., 485) could be enhanced with important information on topics including living arrangements, falls risk, and medication issues such as self-administration capability. In addition, OASIS has categorical assessment measures which could provide rich information for the primary care team, for example, types and degree of functional limitations. Future research would investigate whether communication of additional structured OASIS data for the TOC would be perceived as useful by the primary care team.

The finding that not all information on the TOC mapped to OASIS (e.g., Medications List) indicates OASIS data was a necessary but not sufficient data source for the TOC. This mismatch suggests an opportunity for future research in other standardized data sources which could provide structured data to the TOC.

The overlap of TOC and HHC patient information reveals opportunities for data sharing, especially in this era of value-based care. Furthermore, promoting interoperability is an aspect of CMS’s Value-Based Care program.\textsuperscript{14}

Comparison of the transitions of care interoperability standard (USCDI) with TOC and HHC patient data indicates the USCDI is incomplete in regards to inclusion of HHC information.

The standard maps either fully or partially to a minority (26\%) of HHC classes faxed to primary care. We recommend the USCDI be expanded to include all structured HHC data currently faxed to primary care. Accordingly, we suggest the USCDI, which specifies the Goals class as patient goals, include a class for clinician goals to align with the TOC and 485. We also propose that HHC agencies use the USCDI to electronically transmit data which is currently faxed. However, a barrier to HHC and primary care interoperability is that USCDI use is premised on Health Information Exchanges (HIE) participation, which is hampered by HHC technical and financial resource constraints.

The USCDI mapped either fully or partially to only 5\% of HHC classes captured in the OASIS and not faxed to primary care. Further studies are needed to understand the importance to primary care clinicians of HHC data that mapped to the TOC yet not currently contained in the USCDI. The importance of the HHC data may include informing about timely access to additional services, identification of additional resources that may aid in the care of the patient, or additional accommodations that may be needed given the current mental state of the individual. We recommend USCDI expansion to incorporate structured OASIS data not currently faxed and perceived as important by primary care.

Ingesting HHC information into primary care EHRs as standardized, structured data could enable presentation of the information in the TOC at the right time and allow application of data management tools thereby supporting clinician decision-making. Improved transition of care data interoperability could benefit patients in addressing their individual needs with this patient-centered approach, while benefiting health care organizations financially as they apply these approaches to mitigate hospitalizations and hospital readmissions. Improved transmission from HHC to primary care would enable data analytics: decision support, machine learning, and predictive modeling. These capabilities would support development of additional insights, further informing the data capture process, the interpretations generated from the system, and their subsequent presentation to the clinical care team in addressing patient outcomes. Future research is warranted to assess the feasibility of this recommendation, and the impact on primary care workflow and patient outcomes.

This study has limitations related to the information sources. The TOC guidance used in this analysis contained information topics which tend to be broader than data elements, requiring interpretation by the authors. Future work would incorporate in the analysis TOCs from diverse settings to improve specificity of TOC data elements. Similarly, the USCDI Version 2 contains data elements which are open to author interpretation. For example, Assessment and Plan of Treatment does not specify which discipline is the focus, a consideration as physician information content differs from that of nurses. Also, the 485 analysis was limited to structured HHC data and excluded narrative data. It is possible that the excluded text data could contain information needed for the TOC, as EHRs in HHC contain information in addition to that captured in the OASIS.\textsuperscript{15,16}
We illustrate the gaps in communication that can be addressed through interoperability. Current approaches in data analytics have been applied towards calculating hospital risk for individual patients. By incorporating this information into the primary care EHR, additional insights can be developed, further informing the data capture process, the interpretations generated from the system, and its subsequent presentation to the clinical care team in addressing patient outcomes.

Conclusions

The national move towards value-based care with hospitalizations and readmissions as key quality and cost drivers underscores the importance of interoperability along transitions in care. The important primary care TOC visit during the transition from hospital to home has information deficits despite the rich structured patient data collected in home health care. Implementation of USCDI between HHC and primary care could bridge the information silos. Study findings indicate that addressing the inadequate USCDI coverage of HHC data is warranted.

References

Consolidated Environmental and Social Data Facilitates Neighborhood-Level Health Studies in Philadelphia

Colin Christie, Sherrie Xie, Avantika R. Diwadkar, Rebecca E. Greenblatt, Alexandra Rizaldi, Blanca E. Himes
Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA

Abstract
A wide range of datasets containing geographically distributed measures of the environment and social factors is currently available, and as low-cost sensors and other devices become increasingly used, the volume of these data will continue to grow. Because such factors influence many health outcomes, researchers with varied interests often repeat tasks related to gathering and preparing these data for studies. We created Sensor-based Analysis of Pollution in the Philadelphia Region with Information on Neighborhoods and the Environment (SAPPHIRINE), offered as a web application and R package, to integrate pollution, crime, social disadvantage, and traffic data relevant to investigators, citizen scientists, and policy makers in the Greater Philadelphia Area. SAPPHIRINE’s capabilities include providing interactive maps and customizable data retrieval to aid in the visual identification of pollution and other factor hotspots, as well as hypothesis generation regarding relationships among these factors and health outcomes.

Introduction
Epidemiologic and Electronic Health Record (EHR) data can be enhanced for the study of conditions in which social and economic variables play a prominent role by linking them to sources of external information via person-specific geocodes. For example, air pollution, a well-established global health hazard consisting of particulate matter (PM) and gaseous chemicals that is associated with increased rates of hospitalization for asthma, COPD, pneumonia, and cardiovascular disease as well as with increased all-cause mortality, can be linked to other datasets to understand its influence on health. Similarly, the Area Deprivation Index (ADI), a validated score of socioeconomic deprivation derived from factors including education, employment, income, and housing quality, can be assigned to individual persons in the U.S. via their residential geocodes to obtain estimates of their socioeconomic status. ADI has been correlated with all-cause, cardiovascular, cancer, and childhood mortality. Although area-level information may not adequately substitute for individual-level information in all cases, its use can still capture important relationships.

In the U.S., air pollution levels are monitored by the Environmental Protection Agency (EPA) to ensure state and local compliance with the Clean Air Act. Specifically, six criteria pollutants are monitored: PM10, which are particles with diameter less than 10 μm, PM2.5, or fine particles with diameter less than 2.5 μm, SO2, NO2, O3, and CO. Regulatory monitors operated by the EPA are highly accurate but are stationed at fixed locations with limited spatial coverage. Pollutant concentrations at locations outside of regulatory stations can be estimated using interpolation methods such as inverse-distance-weighting, which can be inaccurate compared to direct measurement. EPA pollution data is also limited in terms of its temporal resolution: measurements are only available as temporal averages over multi-hour periods.

Concern for pollution’s effects on health along with broad demand for accessible environmental monitoring have led researchers and manufacturers to develop low-cost, portable pollution sensors in recent years. These sensors are able to take measurements anywhere and at any time. PM sensors, such as those offered by HabitatMap and PurpleAir, can be used in conjunction with mobile devices to upload PM measurements to crowdsourced databases in real-time. In Philadelphia, HabitatMap and PurpleAir sensors have been utilized by citizen scientists, researchers, and the Clean Air Council, an environmental non-profit, to assess neighborhood-level exposures to air pollution. Other health-related variables, including neighborhood socioeconomic status, crime incidence, and exposure to vehicular traffic, exhibit high geospatial variability in Philadelphia and have been linked to negative health outcomes such as asthma exacerbations. Understanding the drivers of geospatially varying health disparities in Philadelphia thus requires fine-scale, neighborhood-level data on the physical and social environments.

We created Sensor-based Analysis of Pollution in the Philadelphia Region with Information on Neighborhoods and the Environment (SAPPHIRINE), a web application and R package, to integrate sensor-based pollution data, EPA
regulatory monitor data, and other social data on Philadelphia neighborhoods in an interactive, versatile tool that facilitates the conduct of local environmental health studies.

Methods

Data Sources and Processing

Geographic boundaries and study time-period

The Greater Philadelphia Area was defined as the Philadelphia-Reading-Camden Combined Statistical Area, which includes Berks, Bucks, Chester, Delaware, Montgomery, and Philadelphia Counties in Pennsylvania; Atlantic, Burlington, Cape May, Camden, Cumberland, Gloucester, and Salem Counties in New Jersey; Kent and New Castle Counties in Delaware; and Cecil County in Maryland20. A shapefile of this region was created based on a U.S. counties borders dataset available online21. The date range for measures of interest was selected as June 1, 2015 to December 31, 2019.

AirBeam data

HabitatMap offers the portable sensors AirBeam and AirBeam2, which measure PM$_{2.5}$ concentration, temperature, and humidity. AirBeam2 additionally measures PM$_{1}$ and PM$_{10}$ concentrations. Location data corresponding to these measures is recorded via paired smartphone GPS recordings. AirBeam and AirBeam2 data were downloaded from HabitatMap’s website22 as tabular datasets of measurements uploaded by any user within the greater Philadelphia area. To ensure SAPPHIRENE included only mobile and stationary outdoor measures, AirBeam data was flagged if it: 1) spanned a region of less than 0.0001 degree of latitude or longitude, 2) included PM$_{2.5}$ values greater than 150 µg/m$^3$, or 3) contained the strings inside, indoor, or fixed in the title. After manual review of geospatial and temporal plots from these sessions, those determined to be indoor and stationary were excluded. Additional AirBeam data taken by members of our lab between October 23, 2017, and July 26, 2018 were incorporated into the final dataset.

PurpleAir data

PurpleAir offers low-cost stationary sensors that continuously record PM$_{2.5}$, PM$_{1}$, and PM$_{10}$ measurements and automatically upload them to a linked website. We downloaded PurpleAir data23 as 15-minute average measures for all sensors located in the Greater Philadelphia Area. Simultaneous data points from each PurpleAir sensor’s “A” and “B” channels were compared. The two measures were averaged if greater than 30 units and within 10% agreement or if less than 30 and within agreement to 3 units. Otherwise, the data points were excluded.

Sensor data processing

Sensor-based data were processed in R$^{24}$. First, data were summarized in data frames according to measurement type (i.e., temperature, humidity, PM$_{1}$, PM$_{2.5}$, PM$_{10}$). Each data frame included the timestamp, latitude, longitude, measurement value, and sensor ID for each data point. Data frames were then joined by timestamp, latitude, longitude, and sensor ID into one data frame. AirBeam or PurpleAir PM$_{2.5}$ measurements were excluded when a synchronous humidity value of 80% or greater was recorded, as hygroscopic growth of particulate matter occurs at approximately 80% humidity and leads to errant measurements in optical sensors$^{25}$. PM measurements from AirBeam2 sensors were included regardless of humidity because AirBeam2 sensors include a particle drying mechanism to avoid hygroscopic error. AirBeam and AirBeam2 data points were consolidated into 10s averages.

Crime

Tabular data capturing all crime incidents that occurred in Philadelphia during the study period as reported to the Philadelphia Police Department was downloaded from OpenDataPhilly$^{26}$. The data was processed into a data frame that included the timestamp, latitude, and longitude of each incident.

Area Deprivation Index (ADI)

2018 ADI national percentile rankings at the block group level were sourced from the Neighborhood Atlas data portal$^8$ which calculated ADI from the American Community Survey 5-year estimates for 2014-2018 and assigned percentile rankings, with 1 indicating the lowest level of disadvantage and 100 indicating the highest level of disadvantage. The R package censusapi was used to gather the specific GEOIDs of the study area’s block groups to be joined with the ADI data and block group geometries from the US Census Bureau’s 2019 cartographic boundary file. National ADI rank data, geocoded by block group, were then saved as a shapefile.
Traffic

Annual Average Daily Traffic (AADT) is the typical daily traffic volume of a segment of road calculated by dividing the total number of vehicles traveling across a road segment in a year by 365\(^2\). Traffic volume line data containing the AADT of all road segments in Pennsylvania was downloaded from the Pennsylvania Department of Transportation Open Data Portal\(^{28}\) and subsetted to the six Greater Philadelphia Area counties in Pennsylvania (Berks, Bucks, Chester, Delaware, Montgomery, and Philadelphia). The traffic data was not subject to temporal subsetting because the AADT for each road segment was measured over a different time period and was not predicted to have fluctuated substantially given that all measurements were recorded in or after the year 2012. The traffic data was rendered as a raster layer in R by resampling the AADT values for all road segments to a high-resolution grid using the “mean” function in the raster library\(^{29}\).

EPA data

Daily-averaged measurements of PM\(_{2.5}\), PM\(_{10}\), SO\(_2\), NO\(_2\), O\(_3\), and CO concentrations from EPA monitors were downloaded from the EPA Air Data Portal\(^{16}\) for the study time period and the Greater Philadelphia Area. Data describing the EPA monitors was downloaded from the EPA Interactive Map of Air Quality Monitors\(^{30}\). For each pollutant, the measurement data frame was joined to the corresponding monitor data frame by Air Quality System (AQS) site ID and parameter occurrence code. These frames were then joined by date, latitude, longitude, and all site information (AQS ID, name, city, state, start date, and end date).

**SAPPHIRINE Design**

SAPPHIRINE was created using the RStudio R shiny library, which facilitates creation of interactive web apps from R\(^{31}\). The app consists of interactive Leaflet maps\(^{32}\) for various environmental variables plotted using the following approaches: average temperature in °C, average relative humidity as a percentage, average PM\(_1\) concentration in \(\mu g/m^3\), average PM\(_{2.5}\) concentration in \(\mu g/m^3\), average PM\(_{10}\) concentration in \(\mu g/m^3\), average SO\(_2\) concentration in ppb, average NO\(_2\) concentration in ppb, average O\(_3\) concentration in ppb, average CO concentration in ppm, total number of reported crimes, estimated average ADI, and estimated average AADT. For sensor-based measures (temperature, humidity, PM), the average value for each bin is calculated from all available data points falling within the bin, given user-specified parameters. Similarly, the ADIs of all block groups whose geocoordinates fall within each bin are averaged. Crime incidents are summed within each bin. For traffic, the pre-existing raster layer is resampled to fit the user-specified grid of bins using bilinear interpolation, which produces an estimated geospatially averaged AADT value for the region covered by each bin. Measurement densities for sensor-based data types are plotted as log-transformed sums of counts for all available data points for a given measurement type that fall within the bin. Raster layers for EPA data are rendered with inverse-distance-weighted interpolation at 1km\(^2\) resolution using the five nearest stations to each point. Code used to create the app, links to the data sources used, and the R package are available at https://github.com/HimesGroup/sapphirine. The app can be accessed at http://sapphirine.org.

**Using SAPPHIRINE to Explore Potential Relationships Between Environmental Variables**

To illustrate how SAPPHIRINE can be used to explore associations between different geospatial variables, we used the app to visualize the distribution of PM\(_{2.5}\) concentrations and traffic near Interstate 676, a major freeway that travels east-west through downtown Philadelphia. To highlight the app’s unique feature of temporal subsetting, we compared a plot of PM\(_{2.5}\) concentrations generated from all available sensor measurements to a plot generated from measurements taken between 4pm-6pm, produced using app controls to capture approximately the evening rush hour. To illustrate how SAPPHIRINE can help identify pollution hotspots and generate hypotheses, we visually identified regions with greatest EPA PM\(_{2.5}\) measurements and contrasted them with other available environmental measures.

**Results**

**SAPPHIRINE Web App Features**

The user interface is shown in Figure 1. On the “Main Map” tab, a custom map of PM\(_1\), PM\(_{2.5}\), PM\(_{10}\), temperature, humidity, crime, ADI, or traffic can be displayed as a color-scaled raster layer based on the selected variable and user-specified criteria (Figure 1A). Users can select specific sensors from which to display data, dates and times of measurements taken, and a range of latitude and longitude coordinates. Additionally, users can select the resolution of the data displayed by choosing the number of rows and columns by which to partition the selected geographic area. Interactive features of the map include the abilities to navigate and zoom in/out on the map, to open an interactive tool for measuring physical distances on the map, to select whether a map should display measurement values or density of measures for a given variable, and to click on the map to open a popup window that lists all available data types for
the bin corresponding to the clicked location, along with measurement values for all data types. Users can download csv files of 1) the subset of original data points used to create the customized map or 2) the geospatial averages corresponding to the bin values on the map. In the former case, the csv contains measurement and crime data, and in the latter case, all variables, including measurement densities. An alternative view of the data is provided in the “Map Grid” tab which displays maps of multiple variables for their side-by-side comparison, which facilitates hypothesis generation via visual exploration (Figure 1B). Under “EPA Map,” users can select a variable for which to display an interpolated raster layer of EPA data over the specified date range. Users can click on blue dots corresponding to monitor locations to display information about the monitoring site and can click on any location in the raster to retrieve the interpolated pollution estimate. The EPA data used to create the map can be downloaded as a csv file.

**Figure 1.** SAPPHIRINE web app screenshots. A) Main Map displays a single variable at one time. B) Map Grid with PM$_{2.5}$ sensor measurements, crime, ADI, and interpolated EPA PM$_{2.5}$ map layers displayed.

**SAPPHIRINE R Package**

The R package saphirine includes objects and functions allowing for similar data visualization and retrieval directly within R. The comprehensive sensor and crime data frame are included as “local.data”, and the EPA data frame as “EPA.data”. The pre-processed poverty and traffic raster layers are included as “poverty.raster” and “traffic.raster.” “GPA_counties” provides the shapefile of the boundaries of the Greater Philadelphia Area counties. “GPA_counties” is easily subsettable with the function “selectGPACounties”, which allows users to select desired counties by providing their names. For reference, correct county names are accessible in “GPACountyNames”. Similarly, “sensor.list” provides the list of sensors included in the dataset. The package defines a new class of object, “EPA_RasterBrick”, similar to “RasterBrick” with an added slot for corresponding monitor data.

The function “customLocalData” allows users to retrieve subsets of “local.data” according to the geographical region, date range, times of day, variables, and sensors they specify. For the geographical region, users can supply their own shapefile, use a subset of “GPA_counties”, or define a square region using latitude and longitude extrema. By default, all variables are included and all of “GPA_counties” is covered. Similarly, “customEPAData” allows users to retrieve data from the EPA frame, subsetted by date range, region, and variables. “localRaster” accepts a subset of “local.data”, ideally an output of “customLocalData”, and generates a brick of raster layers corresponding to the included variables. Users specify the region of coverage and numbers of rows and columns in each raster as well as the variables to be included in the brick. By default, all sensor variables and crime are included, as are poverty and traffic, which are resampled from “poverty.raster” and “traffic.raster”. Furthermore, users can toggle the addition of measurement density layers for sensor-based variables. “intEPARaster,” similarly, accepts an output of “customEPAData” and generates a raster brick for specified variables over a specified region. The rasters are rendered using inverse-distance-weighted interpolation at 1km$^2$ resolution using the five nearest stations. Users can adjust the inverse distance exponent if desired. Finally, “localMap” and “EPAMap” accept outputs of “localRaster” and “intEPARaster”, respectively, and integrate them into interactive Leaflet maps analogous to those on the web app. Users 1) can choose whether to include descriptive popups, 2) can supply a shapefile for customized display boundaries, and 3) can customize colors for the raster displays.
SAPPHIRINE uses a total 4,355,563 data points, of which 3,606,899 (~82.8%) correspond to environmental variable measures (temperature, humidity, and PM concentrations) collected by low-cost sensors, and 734,616 (~16.9%) are crime incidents reported by the Philadelphia Police Department (Table 1).

**Table 1.** Number of data points used to develop SAPPHIRINE by data source and type.

<table>
<thead>
<tr>
<th></th>
<th>Himes Lab AirBeam Sensors</th>
<th>All AirBeam Sensors</th>
<th>PurpleAir Sensors</th>
<th>EPA Stations</th>
<th>Philadelphia Police Department</th>
<th>Neighborhood Atlas</th>
<th>PennDOT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temperature</td>
<td>77,953</td>
<td>145,712</td>
<td>611,003</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Humidity</td>
<td>79,229</td>
<td>145,091</td>
<td>611,108</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PM1</td>
<td>9,413</td>
<td>586,706</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PM2.5</td>
<td>81,680</td>
<td>149,610</td>
<td>576,724</td>
<td>70,217</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PM10</td>
<td>9,413</td>
<td>561,880</td>
<td>1,968</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SO2</td>
<td></td>
<td></td>
<td>42,766</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NO2</td>
<td></td>
<td></td>
<td>11,980</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>O3</td>
<td></td>
<td></td>
<td>44,741</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CO</td>
<td></td>
<td></td>
<td>28,567</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Crime</td>
<td></td>
<td></td>
<td>734,616</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ADI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5,014</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AADT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>9,034</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Traffic volumes on different road segments are measured sporadically.

**Exploring the Low-Cost Sensor Data at Varying Spatial Resolutions**

Figure 2 demonstrates how users can compare low-cost sensor measurements at different spatial resolutions to understand them better. In Figure 2A, a PM$_{2.5}$ hotspot is apparent in the upper right-hand corner. Figure 2B shows this same region rendered at increased resolution via zooming, which permits interactively determining that the bin concentration is approximately 44 µg/m$^3$ taken over 42 data points. The user can increase the resolution further by adjusting the bin size and spatial coverage, which yields a map such as that displayed in Figure 2C, where it becomes clearer that the hotspot was due to walks along a multi-blocked loop in the Fox Chase neighborhood of Northeast Philadelphia. By changing dates, it becomes apparent that the high measures were taken on a single day: December 3, 2017. Thus, an isolated phenomenon was most likely responsible for a higher PM$_{2.5}$ concentration this particular day.

![Figure 2. Low-Cost Sensor Data at Varying Spatial Resolutions. A) Low-cost sensor measures averaged over full study period. B) Zoomed-in area that has hotspot (boxed region in Figure 2A) with the summary data of one bin displayed. C) high resolution rendering shows that hotspot resulted from a single set of measures along a specific walk.](image-url)
Exploring the Relationship between Low-Cost Sensor PM$_{2.5}$ Measures and Traffic Volume

Figure 3 shows plots produced by SAPPHIRINE of PM$_{2.5}$ concentrations and traffic volumes near I-676. As expected, traffic volume is highest along major freeways (I-676 running east-west, I-76 running north-south in the west, and I-95 running north-south in the east; Figure 3C). While there is little apparent correlation between traffic volume and PM$_{2.5}$ concentrations when PM$_{2.5}$ is estimated from all available sensor data (compare Figure 3B and 3C), such a relationship is patent when PM$_{2.5}$ is estimated from only measurements taken between 4pm and 6pm (compare Figure 3A and 3C). During the evening rush hour, sensor measures of PM$_{2.5}$ recorded near highly trafficked roads tend to appear higher than sensor measures recorded further away (Figure 3A and 3C). This finding may be expected because traffic volume is highest during peak commute times, when highly trafficked roads are most likely to contribute to local pollution levels. PM$_{2.5}$ estimated using all available sensor data include copious measurements recorded during low-traffic times, during which non-traffic sources of pollution are likely to make relatively greater contributions to local PM concentrations. These results confirm the well-established relationship between traffic volume and high PM pollution levels near roadways.$^{33,34}$

![Figure 3](image3.png)

**Figure 3.** Plots of PM$_{2.5}$ concentrations estimated using sensor measures taken A) between 4pm-6pm and B) at all times of day. C) Average traffic volume expressed as average annual average daily traffic (AADT).

Exploring a Regulatory Monitor PM$_{2.5}$ Hotspot versus ADI

Figure 4 shows plots produced by SAPPHIRINE of EPA PM$_{2.5}$ measures and ADI. Visual inspection of the pollution plot shows areas of higher PM$_{2.5}$ concentration, with the blue arrow pointing to Chester, PA, an Environmental Justice area where residents are exposed to pollution from a trash incinerator, a sewage incinerator, chemical plants, and toxic waste sites. On the ADI map, one can readily observe that this region has values near the 100th percentile, indicating high socioeconomic disadvantage. These observations confirm an expected local relationship, demonstrating that SAPPHIRINE can be used to uncover novel hypotheses by linking environmental data types.

![Figure 4](image4.png)

**Figure 4.** Plots of A) interpolated EPA-measured PM$_{2.5}$ concentration and B) ADI. Arrows highlight a PM$_{2.5}$ hotspot in Chester, PA with high socioeconomic disadvantage.
Discussion

We created the SAPPHIRINE web app and R package to integrate geospatially varying environmental and social data in the Greater Philadelphia Area. SAPPHIRINE allows users quickly to identify potential pollution hotspots and explore relationships between environmental and social variables, as well as consider these variables for health studies. Researchers can download data according to their preferences for use in more extensive analyses. Concerned citizens can visualize pollution levels near home or along a commuter route and investigate their source. While there are national- and state-level environmental datasets available, we restricted the geographic region covered by SAPPHIRINE to the Greater Philadelphia Area because local stakeholders lack a resource like it that includes unique local datasets (e.g., crime data) and for which local action can be taken (e.g., deploying low-cost sensors at specific locations). For example, local environmental organizations such as the Clean Air Council can more easily explore the social and environmental issues that impact Philadelphia’s neighborhoods.

SAPPHIRINE is novel in its inclusion of low-cost pollution sensor data and social data alongside more conventional environmental datasets. With the rise in personal sensor use, there has been a concurrent rise in crowdsourced online databases offering visualization tools, but these existing web apps are limited to data generated by specific brands of sensors. Thus, SAPPHIRINE contains a greater quantity and variety of data that can be used to gain insights.

There are limitations to the datasets included in SAPPHIRINE. While EPA monitors measure PM continuously using highly accurate regulatory equipment, low-cost sensors like AirBeam are less reliable, are not as accurate, and are deployed only when researchers or citizen scientists choose to use them. PM estimates derived from portable sensors may be informed by few sessions and not reflect average levels over longer time periods. Fixed-location sensors, such as PurpleAir, tend to be operated continuously, and thus, while they can provide greater temporal resolution, they have limited spatial coverage. While the multiplicity of data layers in SAPPHIRINE provides a unique opportunity for easy hypothesis generation, disparities in data coverage among different variables limit the ability of researchers to address some questions. Low-cost sensor data nonetheless provides high granularity helpful for hypothesis generation and complements data from regulatory monitors.

As advances in technology increase the accuracy and usability of low-cost pollution sensors, we expect a notable increase in crowd-sourced pollution data coverage and reliability. For example, some sensors have been rendered compact enough to attach to a smartphone case, and microchip-sized sensors promise easy auxiliary attachment to personal devices such as smartphones or smartwatches in the future. Such developments will help resolve the data coverage disparities among variables in SAPPHIRINE, thus increasing data interpretability and application utility.

Maintaining SAPPHIRINE up to date with its current data types is straightforward because all data processing scripts are highly automated. In ongoing work, we are updating SAPPHIRINE to include additional environmental datasets including tree cover, green spaces, and neighborhood cohesion. Additionally, we may expand the geographical scope of our application beyond the Greater Philadelphia Area.

Conclusion

We created SAPPHIRINE, an online web application and R package, that integrates geospatially varying environmental and social data in the Greater Philadelphia Area to facilitate their use for research, policy, and citizen-science efforts. Data available in SAPPHIRINE can be customized according to several user-defined parameters and then downloaded, which facilitates its inclusion into EHR-derived and epidemiologic datasets for studies that seek to better understand the role of environmental and social factors on health outcomes.

Acknowledgements

This work was supported in part by National Institutes of Health (NIH) awards R01HL133433, R01HL141992, and P30ES013508.

References


38. Smart Citizen [Internet]. [cited 2021 Mar 10]. Available from: https://smartcitizen.me/kits/


Racial Disparity in Clinical Alert Overrides

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

Abstract

The existence of systemic racism in US health care is widely recognized, but the role that informatics plays has received little attention. Clinical guidelines, which can incorporate implicit racial bias or be adhered to in racially disparate ways, are often the basis for clinical alerting systems. It is also possible that clinicians might be discriminatory in their response to alerts (for example, by deciding whether to agree or override the alert). We sought to study whether alert logic in our hospital uses patient race as part of its criteria and if alert override rates show any racial disparities. We obtained data on 5,120,114 alert events at the University of Alabama at Birmingham (UAB) Hospital and examined override the rates and reasons with respect to patient race. We found override rates of 82.27% and 81.30% for Black or African American patients and White patients, respectively. Some differences by alert were statistically significant but generally small. Override patterns varied by clinician but reasons given were generally not disparate, suggesting that if racist behavior is present, it is not widely systemic. However, the great variability in individual clinician behavior suggests that deeper analysis is warranted to determine whether disparities are indeed racist in nature.

Introduction

The existence of systemic racism in US health care is widely recognized, with increasing study and discussion in the medical literature (see Figure 1). Guidelines for publishing papers discussing health inequities due to racism have been published. Despite the important role that informatics plays in health care systems and structures, relatively little has been published relating informatics to systemic racism. As part of a larger program to achieve anti-racist health informatics, we seek to study ways in which health information systems might abet or mitigate discriminatory practices, including racist, sexist, and ageist, by looking for evidence of disparate activity.

Clinical guidelines, intended to support unbiased, evidence-based care, often treat race as a clinically relevant factor in risk estimation or treatment selection. This inclusion of race may unintentionally result in discriminatory medical practice, for example by steering minority patients away from needed medical procedures. Even when guidelines do not explicitly include racial factors, clinicians with conscious or unconscious (implicit) bias, may apply them in a discriminatory manner, for example by withholding pain medications from Black patients because of beliefs about pain tolerance or tendency towards opiate use disorder.

Electronic health records (EHRs) frequently employ computerized alerts and reminders (collectively referred to in this paper as “alerts”) to help assure adherence with clinical guidelines. Depending on the alerting logic, clinicians may be given an opportunity to choose whether to accept or override the alert. Thus, the same racial disparities that occur with guideline adherence might be replicated by users of health information systems. It is also possible that clinicians might be discriminatory in their response to alerts (for example, by deciding whether to agree or override the alert). Indeed, recent research using machine learning to predict clinician overrides showed some preliminary data suggesting that patient race might correlate with override rate.

Figure 1: Annual publications cited in PubMed, mentioning systemic racism, based on a search of PubMed on March 9, 2021 (https://pubmed.ncbi.nlm.nih.gov/).

Electronic health records (EHRs) frequently employ computerized alerts and reminders (collectively referred to in this paper as “alerts”) to help assure adherence with clinical guidelines. Depending on the alerting logic, clinicians may be given an opportunity to choose whether to accept or override the alert. Thus, the same racial disparities that occur with guideline adherence might be replicated by users of health information systems. It is also possible that clinicians might be discriminatory in their response to alerts (for example, by deciding whether to agree or override the alert). Indeed, recent research using machine learning to predict clinician overrides showed some preliminary data suggesting that patient race might correlate with override rate.

---

1 Related terms, often used interchangeably, are “systematic racism” and “structural racism”.
2 Of the 470 citations found by searching PubMed for “systemic racism” or ‘systematic racism’ or 'structural racism’ on March 7th, 2021, only two include the word “informatics”.
In light of recent raised awareness of the need for proactive antiracism, we sought to study whether alert logic in our hospital uses patient race as part of its criteria and if alert override rates show any racial disparities. Where disparities were noted, we examined individual clinician behavior and the reasons given for overriding alerts.

**Methods**

**Setting**

The University of Alabama at Birmingham (UAB) Hospital is a 1,157-bed tertiary hospital and academic health science center located in Birmingham, Alabama. The hospital has installed Millennium (Cerner Corporation, Kansas City, MO) as its inpatient and outpatient EHR. Millennium provides a capability for systems personnel to define logic for issuing alerts to users when events (such as orders and procedures) do or do not occur. Some alerts simply deliver a message to the user and possibly block the user from carrying out an activity until a preventing state is corrected or removed. Other alerts allow clinicians to provide a reason for overriding the warning message and then allow them to proceed with the activity that triggered the alert.

**Dataset**

We obtained data on all instances of automated alerts issued by the UAB Hospital EHR for the 2019 calendar year for all clinical settings (e.g., in-patient, ICU, clinic, emergency room, etc.). All data were obtained in “limited data set” form with an exemption of oversight from the UAB Institutional Review Board and were subsequently rendered anonymous through removal of date information. Small racial categories were eliminated from analysis. Alerts that did not allow override were excluded. Alerts have obscure names that generally have meaning to system developers; those important to this study have been given more “user-friendly” names by the author. Data included patient race and gender (provided by patients at time of their registration into the health system), a unique (but deidentified) code for the clinicians receiving the alerts, whether the alerts were accepted or overridden and, if overridden, the reason for the override, chosen by the clinician from a standard list. Multiple alerts occur for most patients. Race of the clinicians is not available.

**Analysis**

Total alerts and percent overrides were calculated by race for the total data set and by each alert. Statistical significance of override rate differences were obtained using an “N-1” Chi-squared calculator (MedCalc; https://www.medcalc.org/calc/comparison_of_proportions.php). For those alerts that occurred frequently, reasons given for overrides compared by race. The override rates for frequent alerts were calculated for individual clinicians.

**Results**

**Dataset**

Overall, 169 types of alerts were triggered 5,120,114 times. Of the 35 alerts that allow overrides, 939,472 alert events occurred, of which 764,166 were overridden (81.34%). Eight race classes with relatively few alerts (American Indian or Alaska Native, 1,607; Asian 11,280; Hispanic or Latino, 9,397; Native Hawaiian/Other Pacific Islander 87; Other, 79; Multiple, 1,410; Decline/Refuse 12,857; and blank, 279;) accounted for 3.94% of all alert events and were excluded from further analysis.

Remaining race classes were “Black or African American” (321,055 events; 82.27% override rate) and “White” (581,410 events; 81.30% override rate). The difference between these override rates (0.97%; 95% CI 0.80% to 1.14%; P<0.0001) corresponded to an excess of 3,114 overrides for Black patients over what would be expected if the rate was the same as for White patients. Table 1 shows the comparisons of override rates for each alert. Differences were generally small (less than 6%), especially for alerts that occurred frequently. Figure 2 shows the actual number of alert overrides for Black patients and the expected number for each alert based on the override rate for White patients. Visible differences are apparent for only the most infrequent alerts (Figure 2D).

We selected the two most frequent alerts (200,000 times or more) and three alerts that occurred very frequently (30,000 or more) and had a rate difference greater than 0.05% for further study. Table 2 shows statistics on numbers of unique clinicians and alert events for these five alerts. The percentage of clinicians who overrode all alert events (and thus had no racial disparities) for each of these five alerts ranged from 41.3% to 93.7%. For clinicians whose override rate was less than 100% for all patients (“mixed response”), the numbers of agreements and overrides are shown in Figure 3. Alert event for patients whose gender is “unknown” are therefore excluded. Override patterns vary greatly by clinician, in part due to differences in the populations of patients for whom the alerts are generated and in part due to
the response to the alerts by the clinician. For example, Clinicians 6, 8 and 9 in Figure 3A only had alerts generated for white female patients. In Figure 3B, Clinician 9 agreed with alerts about half the time for all demographic subgroups, while all other clinicians favored overrides 90% of the time for all demographic subgroups.

Override reasons provided by clinicians are shown in Table 3. Racial differences in the rates for reasons given are all below 9%, with the vast majority being below 1%. Despite their small absolute size, many are significantly different statistically.

**Discussion**

Analysis of data from the UAB Hospital’s clinical alerting system show differences in override rates associated with race that while small in absolute terms, do not appear ascribable to chance alone. The data neither support nor exclude the possibility that some of these differences are due to implicit bias or inappropriate discriminatory processes. Individual clinician patient populations and override behavior vary greatly, and the apparent disparities may warrant
further investigation. However, the reasons given for overrides show strong disparities, suggesting that racist behavior may be present. Given the variable nature of the clinical issues addressed by the collection of alerts, statistics alone cannot tell the whole story. An examination of the alert logic itself is also warranted.

The first question to ask, after Vyas and colleagues,[4] is whether alert logic includes consideration of race in a way that might abet explicit or implicit bias. A review of the 35 alerts in this study shows no explicit inclusion of race in the alert logic. However, three alerts related to use of sulfonylurea (Alert 16, from Table 1), radiocontrast dye (Alert 26) and meperidine (Alert 33) do use an estimated glomerular filtration rate (eGFR) calculation that is adjusted for Black race. As noted by Vyas, although such calculations are considered “standard”, they might induce discrimination based on either over-estimation or under-estimation in renal function. In our data, the override rates were slightly

Figure 3: Override counts of five alerts of interests from Table 1 (Graphs A to E – numbers in parentheses correspond to alert numbers in Table 1) for the top 10 most frequent overriding clinicians (numbers 1 to 10) in each alert. Clinician 3 in Graph B is also Clinician 9 in Graph C. Otherwise, clinicians are unique across all five graphs.

3 UAB_EKM!UAB_RULE111_SULF_UREA_ALT, UAB_EKM!UAB_RULE355_RADONC_CREATI and UAB_EKM!UAB_RULE62_MEPERIDINECRL, respectively.
lower for Black patients for one alert (accounting for 7 fewer overrides) and essentially indistinguishable for the other two. If the eGFR calculation is having an effect, it is small, infrequent and, in any case, may still be appropriate.

A second question to ask is whether known discriminatory practices that affect guideline adherence are reflected in alert override activity, such as withholding expensive treatments or delaying decision making. Discriminatory practices related to opioid use, for example, are more than theoretical.[4] Only one of the alerts studied – the meperidine alert (Alert 33) – involves administration of opioid therapy (a warning about the use of meperidine in patients with decreased creatinine clearance). This alert was triggered only 203 times with an override rate difference of less than 0.4%, or less than one fewer overrides for Black patients than would be expected for the same number of alerts for White patients. While the number of alert events is small, it is reassuring to find that alert override behavior does not appear to recapitulate clinician biases seen that have been reported in opioid guideline adherence.

More detailed study of causes for alert override disparities will be difficult, likely involving chart review and interviews to better understand the clinical justification in each case. Information such as clinician race and gender could be collected at that time. The good news is that the racial disparities found, while statistically significant, seem clinically insignificant. The rate differences in the top three alerts in Table 1, correspond to a total of 1,760 excess overrides in Black patients out of a total of 109,931 alerts (1.6%), while the single most frequent alert (Alert 13, “Patient Has Allergy to Drug Being Ordered”[4]) shows 273 fewer overrides (0.2%) than would be expected by chance. Thus, more in-depth study may have difficulty finding sufficient numbers of inappropriate acceptances or overrides for meaningful analyses. And although systemic racial bias should always be a concern, there may very well be excellent reasons for overrides that do have a justifiable correlation with race, for example if decisions differ based on an underlying disease that is itself correlated with race independent of social determinants of health and disease (e.g., ignoring a drug-allergy interaction in a life-threatening situation such as sickle cell crisis).

This study is limited in several ways. First, it involves a single institution – one that is actively involved in addressing systemic racism in healthcare.[9] Because clinician behavior at other institutions may vary, this study should be replicated elsewhere. Second, this study only examined clinical decisions based on alerts implemented with override capabilities. Therefore, it does not attempt to discover racial disparities in other types of decision making or decision support systems. Third, the clinical context of the alert occurrences is not included in this analysis. Alerts occurring in different settings (for example the maternity ward, cancer center, emergency department or medical clinic) may have different patient demographics and may also have different reasons for alert overrides. Fourth, the reasons for override are chosen from a list that may subject to inaccuracies, the list may not contain the desired reason, or the clinician may not choose carefully.[10] Finally, the results of the analysis are only as good as the data themselves. In particular, patient race data in EHRs are frequently inaccurate.[11] Although UAB has special interests in collecting accurate race data[9] and appears to do so based on our anecdotal experience in clinical settings with patients and their EHRs, no formal study has been conducted. If Black patients are being labeled as White by default (that is, in the absence of accurate information), the size of disparities might be being masked, favoring the null hypothesis.

Conclusions

Mitigation of unconscious human bias in clinical decision making requires constant vigilance. Clinical informaticians, being human, must also be vigilant against the transference of their biases into the systems they build. The current study provides a method with the potential for detecting such bias and, at least in this case, suggests that systemic bias is minimal or absent. However, the great variability in individual clinician behavior suggests that deeper analysis is warranted to determine whether disparities that are indeed racist in nature.

Acknowledgments

This study was initiated by discussions with Dr. Tim Kennell, whose work on alert override prediction raised concerns about racial disparity, with Dr. Suzanne Bakken, who encouraged the analysis that informed this paper, and Andria Cimino, who provided editorial and scientific advice. Additional discussions with Drs. Tiffani Brite, Kevin Johnson, Enieda Mendonca, Umberto Tachinardi and Casey Overby Taylor provide additional insight into data analysis. The author is supported in part by the UAB Informatics Institute and by the Center for Clinical and Translational Science, under grant 1TL1TR001418-01 from the National Center for the Advancement of Translational Science (NCATS). The author thanks Ms. Ayme Miles for assistance in obtaining the data used in this study.

4 MUL_MED!DRUGALLERY
References

Table 1: Alert overrides for White and Black patients, by alert type. Five frequent alerts with large, statistically significant racial differences are highlighted in bold font. Details of override reasons for these alerts, along with less cryptic names, are shown in Table 2.

<table>
<thead>
<tr>
<th>Alert Name</th>
<th>Total Alerts</th>
<th>Override Rates</th>
<th>Rate Difference</th>
<th>Confidence Interval</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>White Black</td>
<td>White Black</td>
<td>White Black</td>
<td>Confidence Interval</td>
<td>P</td>
</tr>
<tr>
<td><strong>1</strong> MUL_MED!ALLERGYDRUG</td>
<td>28580 25124</td>
<td>94.34 98.15</td>
<td>3.81</td>
<td>3.50% to 4.13%</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td><strong>2</strong> UAB_EKMIUAB_ADE1_HT_WT_ALLERGY</td>
<td>20685 10584</td>
<td>11.64 17.13</td>
<td>-5.49</td>
<td>-6.34% to -4.66%</td>
<td>0.0001</td>
</tr>
<tr>
<td><strong>3</strong> UAB_EKMIUAB_RULE15_war_INF_SIGN</td>
<td>1196 661</td>
<td>20.15 27.69</td>
<td>7.53</td>
<td>3.50% to 11.7%</td>
<td>0.0002</td>
</tr>
<tr>
<td><strong>4</strong> UAB_EKMIUAB_RULE90_DOPAMINERGIC</td>
<td>279 31</td>
<td>91.40 70.97</td>
<td>-20.43</td>
<td>-38.2% to -6.93%</td>
<td>0.0005</td>
</tr>
<tr>
<td><strong>5</strong> AMB_EKMIAMB_RULE10_FEC_SPEC_ALERT</td>
<td>5413 2320</td>
<td>90.30 87.33</td>
<td>2.97</td>
<td>-4.58% to -1.44%</td>
<td>0.001</td>
</tr>
<tr>
<td><strong>6</strong> UAB_EKMIUAB_ADE6_DIGOXINLAB</td>
<td>1326 707</td>
<td>90.65 87.13</td>
<td>-3.52</td>
<td>-6.56% to -0.692%</td>
<td>0.014</td>
</tr>
<tr>
<td><strong>7</strong> UAB_EKMIUAB_ADE9_ANTIHISTELD_V2</td>
<td>1136 371</td>
<td>92.78 93.79</td>
<td>1.01</td>
<td>0.125% to 1.85%</td>
<td>0.0256</td>
</tr>
<tr>
<td><strong>8</strong> UAB_EKMIUAB_RULE37_PO_CONV_ALERT</td>
<td>36583 26048</td>
<td>95.57 95.92</td>
<td>0.35</td>
<td>0.028% to 0.668%</td>
<td>0.033</td>
</tr>
<tr>
<td><strong>9</strong> UAB_EKMIUAB_RULE184_CHESTAPALERT</td>
<td>42480 22909</td>
<td>24.75 24.02</td>
<td>-0.73</td>
<td>-1.42% to -0.039%</td>
<td>0.0384</td>
</tr>
<tr>
<td><strong>10</strong> MUL_MED!DRUGDRUG</td>
<td>123174 74223</td>
<td>88.34 88.64</td>
<td>0.30</td>
<td>0.009% to 0.589%</td>
<td>0.0434</td>
</tr>
<tr>
<td><strong>11</strong> UAB_EKMIUAB_RULE90_DOPAMIN_MEDM</td>
<td>248 21</td>
<td>90.32 76.19</td>
<td>-14.13</td>
<td>-35.6% to -0.262%</td>
<td>0.0454</td>
</tr>
<tr>
<td><strong>12</strong> UAB_EKMIUAB_ADE2_RCMDMKIDNEY</td>
<td>14997 5804</td>
<td>95.41 94.76</td>
<td>-0.65</td>
<td>-1.33% to -0.005%</td>
<td>0.0486</td>
</tr>
<tr>
<td><strong>13</strong> UAB_EKMIUAB_ADE16_NITRO_ACIDOSIS</td>
<td>113 58</td>
<td>90.27 81.03</td>
<td>-9.23</td>
<td>-21.8% to 1.33%</td>
<td>0.0888</td>
</tr>
<tr>
<td><strong>14</strong> UAB_EKMIUAB_ADE16_NITRO_ACIDOSIS</td>
<td>2787 1776</td>
<td>56.76 59.29</td>
<td>2.53</td>
<td>-0.410% to 5.45%</td>
<td>0.0917</td>
</tr>
<tr>
<td><strong>15</strong> UAB_EKMIUAB_RULE15_war_INF_SIGN</td>
<td>143 61</td>
<td>65.73 54.10</td>
<td>-11.64</td>
<td>-25.0% to 2.75%</td>
<td>0.1174</td>
</tr>
<tr>
<td><strong>16</strong> UAB_EKMIUAB_RULE22_RAD_NM_WT_LMT</td>
<td>48 105</td>
<td>8.33 2.86</td>
<td>-5.48</td>
<td>-16.9% to 1.78%</td>
<td>0.1338</td>
</tr>
<tr>
<td><strong>17</strong> UAB_EKMIUAB_RULE22_RAD_NM_WT_LMT</td>
<td>241 12</td>
<td>90.46 100.00</td>
<td>-9.54</td>
<td>-14.9% to 13.9%</td>
<td>0.2625</td>
</tr>
<tr>
<td><strong>18</strong> UAB_EKMIUAB_RULE35_WAR_INR_SR_AMB</td>
<td>432 283</td>
<td>91.67 93.64</td>
<td>1.97</td>
<td>-2.17% to 5.74%</td>
<td>0.2999</td>
</tr>
<tr>
<td><strong>19</strong> UAB_EKMIUAB_RULE35_WAR_INR_SR_AMB</td>
<td>3 1</td>
<td>66.67 100.00</td>
<td>33.33</td>
<td>-50.5% to 79.2%</td>
<td>0.333</td>
</tr>
<tr>
<td><strong>20</strong> UAB_EKMIUAB_RULE35_WAR_INR_SCRATC</td>
<td>408 333</td>
<td>84.80 82.88</td>
<td>-1.92</td>
<td>-7.35% to 3.36%</td>
<td>0.4789</td>
</tr>
<tr>
<td><strong>21</strong> UAB_EKMIUAB_RULE90_DOPAMINERGIC</td>
<td>426 38</td>
<td>79.81 76.32</td>
<td>-3.50</td>
<td>-19.4% to 7.94%</td>
<td>0.6092</td>
</tr>
<tr>
<td><strong>22</strong> UAB_EKMIUAB_RULE355_RADONC_CREATIVE</td>
<td>739 254</td>
<td>53.32 54.72</td>
<td>1.41</td>
<td>-5.71% to 8.40%</td>
<td>0.6996</td>
</tr>
<tr>
<td><strong>23</strong> UAB_EKMIUAB_RULE379_ED_BUPRENOPH</td>
<td>282 34</td>
<td>87.59 85.29</td>
<td>-2.29</td>
<td>-18.1% to 7.04%</td>
<td>0.7043</td>
</tr>
<tr>
<td><strong>24</strong> UAB_EKMIUAB_RULE268_HALOPERDL65YO</td>
<td>2572 1374</td>
<td>96.11 96.29</td>
<td>0.18</td>
<td>-1.14% to 1.38%</td>
<td>0.7789</td>
</tr>
<tr>
<td><strong>25</strong> UAB_EKMIUAB_RULE222_UABH_C_CT_WTLM</td>
<td>24 33</td>
<td>4.17 3.03</td>
<td>-1.14</td>
<td>-17.4% to 11.6%</td>
<td>0.8195</td>
</tr>
<tr>
<td><strong>26</strong> UAB_EKMIUAB_RULE92_ACE_PHER_ALERT</td>
<td>24 44</td>
<td>91.67 93.18</td>
<td>1.52</td>
<td>-11.4% to 19.6%</td>
<td>0.821</td>
</tr>
<tr>
<td><strong>27</strong> UAB_EKMIUAB_RULE88_FIT_TEST_ALERT</td>
<td>1863 1011</td>
<td>99.73 99.70</td>
<td>-0.03</td>
<td>-0.624% to 0.380%</td>
<td>0.8846</td>
</tr>
<tr>
<td><strong>28</strong> UAB_EKMIUAB_RULE311_ENOX_DOACB</td>
<td>1898 962</td>
<td>43.94 43.76</td>
<td>-0.18</td>
<td>-4.01% to 3.68%</td>
<td>0.927</td>
</tr>
<tr>
<td><strong>29</strong> UAB_EKMIUAB_RULE62_MEPERIDINECRCRL</td>
<td>131 72</td>
<td>65.65 65.28</td>
<td>-0.37</td>
<td>-14.2% to 12.7%</td>
<td>0.9578</td>
</tr>
<tr>
<td><strong>30</strong> UAB_EKMIUAB_RULE257_NAC_TYL</td>
<td>75 14</td>
<td>72.00 71.43</td>
<td>-0.57</td>
<td>-28.1% to 19.8%</td>
<td>0.9654</td>
</tr>
<tr>
<td><strong>31</strong> AMB_EKMIAMB_RULE11_TIMED_UR_ALERT</td>
<td>4540 2644</td>
<td>91.37 90.54</td>
<td>0.82</td>
<td>-0.528% to 2.24%</td>
<td>23.43</td>
</tr>
</tbody>
</table>
Table 2: Clinician behavior for alerts of interest from Table 1. “Mixed response” means that the clinicians did not simply override all events for the alert. “Top 10” are mixed-response clinicians with the most events for the particular alert.

<table>
<thead>
<tr>
<th>Alert</th>
<th>Events</th>
<th>Clinicians</th>
<th>Overrides (%)</th>
<th>Clinicians with Mixed Response (%)</th>
<th>Events for Top Ten Clinicians</th>
<th>Overrides for Top 10Clinicians (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Allergy for Drug Patient Is Currently Taking</td>
<td>53443</td>
<td>1414</td>
<td>51622 (96.6)</td>
<td>195 (13.8)</td>
<td>997</td>
<td>875 (87.8%)</td>
</tr>
<tr>
<td>2. Patient Allergy Height or Weight Undocumented</td>
<td>7812</td>
<td>2758</td>
<td>4220 (54.0)</td>
<td>257 (9.3)</td>
<td>1999</td>
<td>1626 (81.3%)</td>
</tr>
<tr>
<td>9. Patient Has Had Recent AP Chest X-ray</td>
<td>54301</td>
<td>2137</td>
<td>16015 (29.5)</td>
<td>1254 (58.7)</td>
<td>10108</td>
<td>9846 (97.4%)</td>
</tr>
<tr>
<td>10. Drug-Drug Interaction</td>
<td>196632</td>
<td>3327</td>
<td>174601 (88.8)</td>
<td>208 (6.3)</td>
<td>15414</td>
<td>11798 (76.5%)</td>
</tr>
<tr>
<td>13. Patient Has Allergy to Drug being Ordered</td>
<td>415408</td>
<td>4045</td>
<td>365532 (88.0)</td>
<td>1075 (26.7)</td>
<td>32187</td>
<td>24551 (72.3%)</td>
</tr>
</tbody>
</table>
Table 3: Override Reasons Given by Clinicians, Based on Race, for Alerts with High Rates of Occurrence and Racially Disparate Override Rates (numbered alert names correspond to numbers from Table 1). Chi-Square values of <0.05, <0.005 and <0.001 are noted with “*”, “**” and “***”, respectively.

<table>
<thead>
<tr>
<th>Reason</th>
<th>White Count</th>
<th>White Rate</th>
<th>Black Count</th>
<th>Black Rate</th>
<th>Rate Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1. MUL_MED!ALLERGYDRUG</strong> – Allergy Reported for Drug Patient is Currently Taking</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Allergy information not correct</td>
<td>101</td>
<td>0.37</td>
<td>37</td>
<td>0.15</td>
<td>-0.22 ***</td>
</tr>
<tr>
<td>Chart Review: No Action</td>
<td>18804</td>
<td>69.74</td>
<td>19219</td>
<td>77.94</td>
<td>8.20 ***</td>
</tr>
<tr>
<td>Contraindicated</td>
<td>30</td>
<td>0.11</td>
<td>18</td>
<td>0.07</td>
<td>-0.04</td>
</tr>
<tr>
<td>Data error</td>
<td>31</td>
<td>0.11</td>
<td>23</td>
<td>0.09</td>
<td>-0.02</td>
</tr>
<tr>
<td>Defer to primary physician</td>
<td>470</td>
<td>1.74</td>
<td>367</td>
<td>1.49</td>
<td>-0.25 *</td>
</tr>
<tr>
<td>Deferring to other priorities</td>
<td>8</td>
<td>0.03</td>
<td>10</td>
<td>0.04</td>
<td>0.01</td>
</tr>
<tr>
<td>Disagree with recommendation</td>
<td>9</td>
<td>0.03</td>
<td>5</td>
<td>0.02</td>
<td>-0.01</td>
</tr>
<tr>
<td>Essential therapy, will take precautions</td>
<td>1377</td>
<td>5.11</td>
<td>676</td>
<td>2.74</td>
<td>-2.37 ***</td>
</tr>
<tr>
<td>Interaction noted, will take precautions</td>
<td>4274</td>
<td>15.85</td>
<td>3084</td>
<td>12.51</td>
<td>-3.34 ***</td>
</tr>
<tr>
<td>Not Applicable - Alabama Organ Center</td>
<td>51</td>
<td>0.19</td>
<td>31</td>
<td>0.13</td>
<td>-0.06</td>
</tr>
<tr>
<td>Not applicable</td>
<td>678</td>
<td>2.51</td>
<td>695</td>
<td>2.82</td>
<td>0.31 *</td>
</tr>
<tr>
<td>Order already exists</td>
<td>202</td>
<td>0.75</td>
<td>79</td>
<td>0.32</td>
<td>-0.43 ***</td>
</tr>
<tr>
<td>Order this agent, will stop other drug</td>
<td>60</td>
<td>0.22</td>
<td>35</td>
<td>0.14</td>
<td>-0.08 *</td>
</tr>
<tr>
<td>PRN Medication</td>
<td>23</td>
<td>0.09</td>
<td>6</td>
<td>0.02</td>
<td>-0.07 **</td>
</tr>
<tr>
<td>Patient allergic</td>
<td>484</td>
<td>1.80</td>
<td>251</td>
<td>1.02</td>
<td>-0.78 ***</td>
</tr>
<tr>
<td>Patient not available</td>
<td>15</td>
<td>0.06</td>
<td>2</td>
<td>0.01</td>
<td>-0.05 **</td>
</tr>
<tr>
<td>Patient refused</td>
<td>29</td>
<td>0.11</td>
<td>5</td>
<td>0.02</td>
<td>-0.09 ***</td>
</tr>
<tr>
<td>Poor results on prior application</td>
<td>3</td>
<td>0.01</td>
<td>2</td>
<td>0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>Reaction does not preclude therapy</td>
<td>254</td>
<td>0.94</td>
<td>100</td>
<td>0.41</td>
<td>-0.53 ***</td>
</tr>
<tr>
<td>Specimen Obtained</td>
<td>3</td>
<td>0.01</td>
<td>8</td>
<td>0.03</td>
<td>0.02</td>
</tr>
<tr>
<td>Treatment plan requirement</td>
<td>54</td>
<td>0.20</td>
<td>5</td>
<td>0.02</td>
<td>-0.18 ***</td>
</tr>
<tr>
<td>mCDS_Filtering_RX</td>
<td>0</td>
<td>0.00</td>
<td>2</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>mCSD_Filtering_IP</td>
<td>2</td>
<td>0.01</td>
<td>0</td>
<td>0.00</td>
<td>-0.01</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>26962</td>
<td></td>
<td>24660</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| **2. UAB_EKMIUAB_ADE1_HT_WT_ALLERGY** – Patient allergies, height or weight need to be documented |             |            |             |            |                 |
| Disagree with recommendation                                          | 215         | 22.26      | 99          | 14.20      | -8.06 ***       |
| Not applicable                                                        | 122         | 12.63      | 134         | 19.23      | 6.60 **         |
| Reaction does not preclude therapy                                     | 40          | 4.14       | 24          | 3.44       | -0.70           |
| Treatment plan requirement                                            | 589         | 60.97      | 440         | 63.13      | 2.14            |
| **Total**                                                             | 966         |            | 697         |            |                 |

| **9. UAB_EKMIUAB_RULE184_CHESTAPALERT** - Patient has had recent (past 3 days) AP chest x-ray |             |            |             |            |                 |
| Disagree with recommendation                                          | 10513       | 100.00     | 5502        | 100.00     | 0.00          |
| **Total**                                                             | 10513       |            | 5502        |            |                 |

| **10. MUL_MED!DRUGDRUG** - Interaction between drug patient is taking and drug being ordered |             |            |             |            |                 |
| Allergy information not correct                                       | 252         | 0.23       | 133         | 0.20       | -0.03          |
| Chart Review: No Action                                               | 16590       | 15.25      | 10132       | 15.40      | 0.15           |
| Contraindicated                                                       | 51          | 0.05       | 43          | 0.07       | 0.02           |
| Data error                                                            | 140         | 0.13       | 116         | 0.18       | 0.05 **        |
| Defer to primary physician                                           | 1209        | 1.11       | 594         | 0.90       | -0.21 ***      |
| Deferring to other priorities                                         | 192         | 0.18       | 163         | 0.25       | 0.07 **        |
| Disagree with recommendation                                          | 359         | 0.33       | 222         | 0.34       | 0.01           |
| Essential therapy, will take precautions | 25252 | 23.21 | 15917 | 24.19 | 0.98 *** |
| Interaction noted, will take precautions | 50670 | 46.57 | 30775 | 46.78 | 0.21 |
| Not Applicable - Alabama Organ Center | 165 | 0.15 | 96 | 0.15 | 0.00 |
| Not applicable | 1156 | 1.06 | 879 | 1.34 | 0.28 *** |
| Order already exists | 2087 | 1.92 | 1151 | 1.75 | -0.17 * |
| Order this agent, will stop other drug | 5503 | 5.06 | 3100 | 4.71 | -0.35 ** |
| PRN Medication | 265 | 0.24 | 146 | 0.22 | -0.02 |
| Patient allergic | 17 | 0.02 | 0 | 0.00 | -0.02 |
| Patient not available | 4 | 0.00 | 13 | 0.02 | 0.02 ** |
| Patient refused | 14 | 0.01 | 5 | 0.01 | 0.00 |
| Poor results on prior application | 14 | 0.01 | 2 | 0.00 | -0.01 * |
| Reaction does not preclude therapy | 3172 | 2.92 | 1500 | 2.28 | -0.64 *** |
| Specimen Obtained | 107 | 0.10 | 38 | 0.06 | -0.04 * |
| Treatment plan requirement | 1572 | 1.44 | 747 | 1.14 | -0.30 *** |
| mCDS_Filtering_HXIP | 5 | 0.00 | 5 | 0.01 | -0.01 ** |
| mCDS_Filtering_IPRX | 5 | 0.00 | 4 | 0.01 | -0.01 ** |
| mCDS_Filtering_RX | 4 | 0.00 | 4 | 0.01 | -0.01 ** |
| mCSD_Filtering_IP | 7 | 0.01 | 4 | 0.01 | 0.00 |
| Total | 108812 | 65789 |

### 13. MUL_MED!DRUGALLERGY – Patient Has Allergy to Drug Being Ordered

| Allergy information not correct | 1397 | 0.57 | 729 | 0.60 | -0.03 |
| Chart Review: No Action | 43184 | 17.71 | 22189 | 18.24 | -0.53 *** |
| Contraindicated | 134 | 0.05 | 50 | 0.04 | 0.01 |
| Data error | 248 | 0.10 | 83 | 0.07 | 0.03 ** |
| Defer to primary physician | 2609 | 1.07 | 1143 | 0.94 | 0.13 ** |
| Deferring to other priorities | 361 | 0.15 | 202 | 0.17 | -0.02 |
| Disagree with recommendation | 675 | 0.28 | 328 | 0.27 | 0.01 |
| Essential therapy, will take precautions | 42792 | 17.55 | 23020 | 18.92 | -1.37 *** |
| Interaction noted, will take precautions | 122119 | 50.08 | 58661 | 48.21 | 1.86 *** |
| mCDS_Filtering_HXIP | 40 | 0.02 | 16 | 0.01 | 0.01 * |
| mCDS_Filtering_IPRX | 108 | 0.04 | 91 | 0.07 | -0.03 *** |
| mCDS_Filtering_RX | 117 | 0.05 | 60 | 0.05 | 0.00 |
| mCSD Filtering_IP | 97 | 0.04 | 48 | 0.04 | 0.00 |
| Not applicable | 3285 | 1.35 | 1576 | 1.30 | 0.05 |
| Not Applicable - Alabama Organ Center | 340 | 0.14 | 126 | 0.10 | 0.04 ** |
| Order already exists | 6648 | 2.73 | 4068 | 3.34 | -0.62 *** |
| Order this agent, will stop other drug | 746 | 0.31 | 387 | 0.32 | -0.01 |
| Patient allergic | 105 | 0.04 | 45 | 0.04 | 0.01 |
| Patient not available | 55 | 0.02 | 32 | 0.03 | -0.01 |
| Patient refused | 114 | 0.05 | 65 | 0.05 | 0.00 |
| Poor results on prior application | 43 | 0.02 | 8 | 0.01 | 0.01 * |
| PRN Medication | 708 | 0.29 | 333 | 0.27 | 0.02 |
| Reaction does not preclude therapy | 16232 | 6.66 | 7546 | 6.20 | 0.45 *** |
| Specimen Obtained | 201 | 0.08 | 49 | 0.04 | 0.04 *** |
| Treatment plan requirement | 1506 | 0.62 | 813 | 0.67 | -0.05 |
| Total | 243864 | 121668 |
Data-Driven Sequential Uptake Pattern Discovery for Family Planning Studies

Celia Cintas, PhD¹, Victor Akinwande, MS¹, Ramya Raghavendra, PhD², Girmaw Abebe Tadesse, PhD¹, Aisha Walcott-Bryant, PhD¹, Charity Wayua, PhD³, Fredrick Makumbi, MHS, Bstat, PhD³, Rhoda K. Wanyenze, MBChB, MPH, PhD⁴ and Komminist Weldemariam, PhD¹

¹IBM Research Africa, Nairobi, Kenya; ²IBM T. J. Watson Research Center, Yorktown, NY, USA; ³Department of Epidemiology & Biostatistics School of Public, College of Health Sciences, Makerere University, Kampala, Uganda; ⁴Department of Disease Control and Environmental Health, School of Public Health, Makerere University, Kampala, Uganda.

Abstract Family planning is a crucial component of sustainable global development and is essential for achieving universal health coverage. Specifically, contraceptive use improves the health of women and children in several ways, including reducing maternal mortality risks, increasing child survival rates through birth spacing, and improving the nutritional status of both mother and children. This paper presents a data-driven approach to study the dynamics of contraceptive use and discontinuation in Sub-Saharan African (SSA) countries. We aim to provide policymakers with discriminating contraceptive use patterns under different discontinuation reasons, contraceptive uptake distributions, and transition information across contraceptive types. We used Demographic Health Survey (DHS) Calendar data from five SSA countries. One recurrent pattern found was that continuous usage of injectables resulted in discontinuation due to health concerns in four out of five countries studied. This type of temporal analysis can aid intervention development to support sustainable development goals in Family Planning.

Introduction

The World Health Organization (WHO) defines Family Planning (FP) as a means to allow individuals and couples to anticipate and attain their desired number of children and the spacing and timing of their births. The effective use of contraceptives can significantly improve the health of women and children, and it has been shown to reduce maternal mortality risks, and improve child survival through birth spacing, as well as improve the nutritional status of both mother and child¹. Contraceptive discontinuation (CD) occurs when the use of a contraceptive is stopped for any reason while the woman is still at risk of an unintended pregnancy. It is estimated that between 20-40% of users discontinue a contraceptive method for reasons other than a desire to become pregnant or no longer needing a method (e.g., post-menopausal)². While a portion of women would switch to a different contraceptive method after discontinuation, many do not, leaving them at risk of an unintended pregnancy. Thus, CD may signal a dissatisfaction with the contraceptive used, and hence it remains an important topic for domain-experts and policy makers in order to improve the efficiency of FP³–⁷. While there are multiple studies on FP and contraception, focusing on measuring fertility rates⁸, childbirth spacing⁹, modelling prevalence of modern methods¹⁰, and grouping of calendar episodes by long and short-term contraceptive methods¹¹, not many studies were conducted to identify temporal patterns that lead to discontinuation over time. Our work aims to uncover temporal patterns of contraceptive discontinuation and reasons for discontinuation across countries and subpopulations of women. Such patterns can lead to further research questions regarding the demand and supply chain issues of contraceptives or how to develop robust educational interventions to provide information regarding contraceptive methods and their side-effects.

This work was done in partnership with the Bill & Melinda Gates Foundation (BMGF), specifically with their Family Planning team who advised on research questions of interest, data sources, and provided feedback on proposed methods and evaluation of results. In this paper, we explore two questions regarding contraception use and discontinuation episodes¹: (Q1) Are there any recurrent sequences of contraceptive use and discontinuation across countries? (Q2) What do women transition to when they discontinue or switch between contraceptive methods? To address these questions, a series of events are analyzed in a time window of 12 months using contraceptive calendar data from the Demographic and Health Surveys Program (DHS)¹². Specifically, in each month, DHS provides 1) the type of contraceptive method being used, if any, 2) the events that occurred in that month such as pregnancy, birth, termination, and

¹A discontinuation episode refers to an event that occurs in a specific month when a contraceptive method is not used for any reason.
3) the reason for discontinuation (if it occurs). While the DHS data provided longitudinal calendar data (for five years) on women’s contraceptive use, the latest 12 months were used to reduce the impact of recall bias. Pattern mining was performed to check for common patterns of events among women that discontinue their contraceptive uptake. In this paper, two major scientific contributions are provided in line with addressing the above research questions. First, discriminatory subsequences are detected and characterized, along with the extent of their discriminatory power, for different types of discontinuation reasons (Q1). Second, we provide an analytic tool to visualize uptake distributions and transition between contraceptive methods in various subpopulations (Q2). The tool enables domain experts to easily visualize contraceptive switching patterns and interact with the obtained results. For the first time, this allows us to automatically compare subgroups of interest using patterns that are unique to a subgroup. Such patterns are valuable to a practitioner to gain a notion into contraceptive uptake time-based patterns and discontinuation trends. The analysis shows that patterns of continuous usage of injectables were a precursor to contraceptive discontinuation due to health concerns in four out of five Sub-Sahara African (SSA) countries studied. Additionally, we observed that such sequences for DHS Kenya (2014), DHS Ghana (2014), DHS Nigeria (2013) and DHS Burkina Faso (2010) are discriminant i.e. statistically unique to women who used injectables.

To the best of our knowledge, this is the first presentation of data-driven and automatic subsequence mining used to understand patterns of contraceptive use in various subpopulations. Furthermore, the proposed dashboard-based interactive and visual analytic tool enables policymakers to explore longitudinal contraceptive dynamics and discontinuation reason via its interface that allows interactive exploration of contraceptive use, distribution, and pattern uptake trends across different countries.

Related Work

Previous studies in this field of research mainly used traditional models, such as decision trees and logistic regressions, to perform stratified analysis to understand the correlations between contraceptive uptake and pre-determined covariates \(^3\text{-}^\text{7,13}\). The majority of these studies are focused on India (the second most populous Asian country) \(^7\) and Nigeria (the most populous African country) \(^3\text{-}^\text{5}\). A few studies were also able to understand the determinants of contraceptive use across multiple countries, particularly Sub-Saharan African (SSA) countries \(^5\text{-}^\text{14}\) and Cahill et al. \(^6\) analysed the prevalence rate across multiple countries in the world. A summary table is provided in Table 1.

Table 1: Summary of related works on contraceptive use. Countries: Kenya (KE), Malawi (MA), Madagascar (MD), Rwanda (RW), Cameroon (CA), Chad (CH), Zambia (ZA), Zimbabwe (ZI), Ghana (GH) and Mali (MI). LARC: long-acting reversible contraceptive; LR: Logistic Regression. Data source: Demographic and Health Surveys (DHS).

<table>
<thead>
<tr>
<th>Authors</th>
<th>Country</th>
<th>Data</th>
<th>Research goal</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Azukie et al. (^3) (2017)</td>
<td>Nigeria</td>
<td>DHS</td>
<td>Identifying predictors of discontinuation</td>
<td>Binomial LR</td>
</tr>
<tr>
<td>Michael (^4) (2017)</td>
<td>Nigeria</td>
<td>DHS</td>
<td>Influence of socio-cultural and economic factors on contraceptive use and desire for less children</td>
<td>Multivariate LR</td>
</tr>
<tr>
<td>Larsson &amp; Stanfors (^13) (2014)</td>
<td>GH, KE, MA, ZA</td>
<td>DHS</td>
<td>The influence of women education and empowerment on contraceptive use</td>
<td>Bivariate and Multivariate LRs</td>
</tr>
<tr>
<td>Adedini et al. (^14) (2019)</td>
<td>MA, Rw, CA, CH, ZA, ZI, GH, MI</td>
<td>DHS</td>
<td>Identifying patterns of LARC</td>
<td>Multinomial LR</td>
</tr>
<tr>
<td>Ours</td>
<td>KE, NG, BF, GH, ET</td>
<td>DHS</td>
<td>Mining sequential episodes from calendar data to understand contraceptive uptake patterns and discontinuation reasons in different subpopulations.</td>
<td>Discriminatory Subsequence Mining and Interactive Visualization Tools</td>
</tr>
</tbody>
</table>
Figure 1: Overview of the proposed framework that takes temporal data, such as contraceptive calendar data available in DHS datasets, for a given country $C_n$. There are two main modules that feed the analytics of the dashboard in the framework. First, the Discriminatory pattern mining module is in charge of extracting the top $k$ discriminatory subsequences between two classes defined by discontinuation reason $R_{m}$ to display on the dashboard. Second, the transition extractor handles preprocessing episode data and probability distribution uptake estimation. This module returns a pair of a probability ($\text{Prob}_n^i$), transition ($\text{Transitions}_n^i$) for each contraceptive method $CM_k$ for $C_n$.

Most existing works on the analysis of FP programs and identifying key determinants of contraceptive use tend to employ simple modeling techniques, such as regressions. These studies largely ignored the time-based calendar data, which we believe may reveal novel patterns on contraceptive use in different groups. In contrast to these prior studies, we employ sequence mining methods to provide data-driven patterns into contraceptive discontinuation, using the sequential calendar data to mine uptake patterns.

Materials and Methods

In this section, the dataset used for the analyses and the extension of PrefixSpan to analyze contraceptive discontinuation is described. An overview of the proposed approach is shown in Figure 1.

The dataset

The DHS program\(^2\) has collected, analyzed and disseminated representative data on population, health, HIV and nutrition through more than 300 surveys in over 90 different countries. Nationally representative surveys are designed to collect data on monitoring and impact evaluation indicators important for individual countries and for cross-country comparisons. The DHS calendar\(^12\) dataset contains monthly history of certain key events in the life of the respondent. The calendar collects a complete history of women’s reproduction and contraceptive use for a period of between five and seven years prior to the survey. The exact length of the period covered by the contraceptive calendar varies depending on the duration of data collection, whether the survey overlapped two years and the month in which the respondent was interviewed\(^12\). In the current standard DHS-7 questionnaire for the calendar consists of two values per episode. The first value represent events such as: births, pregnancies, terminations and contraceptive use (twenty six different types of episodes are considered). The second value represents the reason for discontinuation of contraceptive use (nineteen different reasons are considered). For this study, 95,855 calendars were processed, spanning five SSA

\(^2\)https://www.dhsprogram.com/Data/

Data-driven Discriminatory Sequence Mining

Sequential pattern mining is a data mining technique that discovers frequent subsequences in a sequence database, such as DHS Contraceptive Calendar data. PrefixSpan\textsuperscript{15} is one of such pattern mining techniques and it employs a projection-based, sequential pattern-growth approach for efficient mining of temporal patterns is used. In this approach, a sequence database is recursively projected onto a set of smaller projected databases, and sequential patterns are grown in each projected database by exploring locally frequent fragments only. Several notations for the contraceptive pattern mining problem being investigated are defined below.

Definition 1

Let \( B = \{ e_1, e_2, \ldots, e_m \} \) be an item base consisting of items, in this case, different types of contraceptive episodes. An itemset is a subset of items if \( I \subseteq B \). A sequence is an ordered list of itemsets, in our case a series of monthly ordered episodes from calendar data. A sequence \( T = (t_1 \rightarrow t_2 \rightarrow \cdots \rightarrow t_n) \) with \( \forall k 1 \leq k \leq n : t_k \subseteq B \).

Definition 2

Let \( I \subseteq B \) be an itemset and \( T \) a sequence over \( B \), the cover of \( I \) is defined as the set:

\[
K_T(I) = k \in 1, \cdots, n | I \subseteq t_k
\]

The cover of an itemset is the index set of sequences that contain all items in \( I \). The value \( \sigma_T = \frac{1}{n} |K_T(I)| \) is called relative support of \( I \). The support of \( I \) is the fraction of sequences that contain it.

Definition 3

Given the minimum support \( \sigma_{\text{min}} \in \mathbb{R}, 0 < \sigma_{\text{min}} \leq 1 \), the set of frequent itemsets is defined as:

\[
\Phi_T(\sigma_{\text{min}}) = I \subseteq B | \sigma_T(I) \geq \sigma_{\text{min}}
\]

Definition 4

A sequence \( s = (s_1, s_2, \cdots, s_n) \) matches a sequence \( s' = (s'_1, s'_2, \cdots, s'_m) \) if there exists \( j_1 < j_2 < \cdots < j_n \) such that \( s_i = s'_j \), the function is defined as \( \text{match}(s, s') \).

In this study, the goal of using sequence mining is to identify contraceptive uptake patterns that may be unique to one cohort of women (e.g., women that discontinued due to health concerns). Extracting such sequences do not only shed information regarding uptake patterns to domain experts, but can also have predictive power —i.e., be able to predict that a discontinuation may occur in \( k \) steps if a particular pattern is being observed. To identify such unique or “discriminatory” sequences, the PrefixSpan\textsuperscript{15} is extended in several ways. First, we need to look for differentiating sequences between two classes, namely women who discontinued due to health concerns and those who did not discontinue. To address this, PrefixSpan was extended to mine for patterns that are different between the two classes of interest. There are two parameters that need to be set, namely \( \sigma_{\text{min}} \) (Def. 3) and \( w_{\text{max}} \). The first parameter expresses a threshold on how often a particular pattern occurs, in terms of the minimum percentage across all the sequences in DHS calendar data for a given country. The patterns identified are sequential in nature and are subsequences of the contraceptive uptake patterns. \( w_{\text{max}} \) defines the time window to be analyzed. PrefixSpan reporting method is extended such that each pattern is tagged with various metrics. While these are familiar concepts in association rule mining, their definitions to mine sequences with discriminatory power are slightly modified in this work. The two key metrics are support and lift. Support left and right, is the support \( \sigma_T \) (Def 2) of a pattern \( T \), for left and right datasets, respectively. Lift(\( T, left, right \)) is the ratio of \( \frac{Pr(\text{match}(s,S_1))}{Pr(\text{match}(s,S_2))} \), defining how often a pattern occurs in the left dataset compared to the right. When this number is very large, it indicates that the pattern is unique to the left dataset. With these definitions, the problem of finding discriminatory patterns can be formalised as follows: Let \( p_s = \frac{Pr(\text{match}(s,S_1))}{Pr(\text{match}(s,S_2))} \) the ratio of finding a given pattern, where \( S_1 \) and \( S_2 \) are two collections of sequences (e.g., those that continue contraceptive use vs. those that don’t). Let \( d_s = \max(p_s, \frac{1}{p_s}) \) denote the discriminatory power of sequence \( s \) with respect to the two collections of sequences \( S_1 \) and \( S_2 \), where a larger \( d_s \geq s \) is a more powerful discriminator.
Given the two collections of sequences $S_1$ and $S_2$, learn as a collection of sequences $S(\|S\| < |S_1|, |S_2|)$ such that $\forall s \in S, d_s$ with respect to $S_1$ and $S_2$ is greater than a given threshold ($th$). We note that pruneByDominance is a function to eliminate those patterns $q$ that are obtained by augmenting an existing pattern $p$, where $p$ is shorter or more general than $q$, and has a higher confidence of predicting a class than $q^{16}$. Algorithm 1 shows the pseudo-code for discriminatory sequence mining (DSM).

**Algorithm 1:** A pseudo-code for the proposed Discriminatory Sequence Mining.

```
\begin{algorithm}
\begin{algorithmic}
\State \textbf{input} : Collections $S_n = [S_1, S_2]$, $\sigma_{\text{min}}$, $I$
\State \textbf{output}: $S_{\text{out}}$, where ($|S_{\text{out}}| < |S_1|, |S_2|$) such that $\forall s \in S, d_s > th$
\For {$c \rightarrow 1$ to $|S_n|$ do}
\State $freq_s = \Phi_s(\sigma_{\text{min}})$;
\State $c_s = \frac{1}{n}Kfreq_x(1)$;
\EndFor
\For {$c \rightarrow 1$ to $|freq_n|$ do}
\If {$S_1$ is Prefix($S_2$) and Lift($freq_s, S_1, S_2$) > Lift($freq_s, S_2, S_1$)}
\State $S_{\text{out}} = \text{pruneByDominance}(freq_s, S_1)$;
\EndIf
\If {match($S_1, S_2$) > th}
\State $S_{\text{out}} = \text{pruneByShadows}(S_1)$;
\EndIf
\EndFor
\end{algorithmic}
\end{algorithm}
```

**Experimental Analysis and Results**

In this section, we present a set of experiments conducted to answer the two research questions outlined in the Introduction, by analysing contraceptive use and discontinuation from different angles to provide a holistic view of the problem landscape to domain experts. First, motivated by (Q1), the discriminatory sequence mining algorithm is used to mine frequent temporal patterns of contraceptive use across all countries. Second, regarding (Q2), contraceptive transitions is studied to answer the following questions: how long an individual continues to use a contraceptive option? And what they transition to when a method is discontinued?

**Experimental setup for Q1: Discriminatory subsequence mining for fine-grained pattern extraction**

**Data Preprocessing**

For sequence mining experiments, more than 95,855 contraceptive calendars corresponding to women population from five Sub-Saharan countries were analysed, i.e., DHS Kenya (2014), DHS Nigeria (2013), DHS Ghana (2014), DHS Ethiopia (2016), and DHS Burkina Faso (2010). Retrospective reporting of contraceptive use and discontinuation rely heavily on the ability of respondents to accurately recall events. To reduce recall bias, the 12 months prior to a discontinuation event are used in this analysis, and the three months immediately before the survey is removed to account for under-reporting of first trimester pregnancies at the time of the survey. For all the experiments presented in this section, calendar data from individuals who wanted to become pregnant are excluded (DHS Code 2).

As described in the Methods section, two parameters are needed to train the Discriminatory Sequence Model, namely the time window to be analyzed ($w_{\text{max}}$) and the minimal support required ($\sigma_{\text{min}}$). These parameters were tuned in consultation with domain experts, understanding that smaller minimal support within a time window provides more sequences needing to be examined for interpretation, and a shorter window provides patterns including short-term contraceptive methods. A time window of $w_{\text{max}} = 12$ months is constructed to use the full calendar year extracted and capture both long- and short-term contraceptive methods. We selected 0.3 as a minimum fraction of the population required for the patterns to be significant for minimal support. In order to work with DSM, two groups for evaluation were determined. For example, Table 2 shows the discrimination patterns of calendar data that contain discontinuations due to health concerns as opposed to the rest of the population. The reason for discontinuation is extracted from the
Table 2: Discriminatory subsequences mined in the subpopulation of women who discontinue due to health concerns (DHS Discontinuation Code 4), as opposed to the rest of the population that discontinue for any other reason. Injectables (DHS Code 3), non-use (DHS Code 0)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>KE</td>
<td>3 → 0</td>
<td>0.33125</td>
<td>0.00909</td>
<td>36.429</td>
</tr>
<tr>
<td>NG</td>
<td>3 → 3 → 3 → 0 → 0</td>
<td>0.33644</td>
<td>0.00005</td>
<td>561.28</td>
</tr>
<tr>
<td>NG</td>
<td>3 → 3 → 0 → 0</td>
<td>0.34579</td>
<td>0.00005</td>
<td>663.40</td>
</tr>
<tr>
<td>GH</td>
<td>3 → 3 → 0 → 0 → 0</td>
<td>0.33888</td>
<td>0.10357</td>
<td>3.2715</td>
</tr>
<tr>
<td>BF</td>
<td>3 → 3 → 0 → 0 → 0 → 0</td>
<td>0.31896</td>
<td>0.14345</td>
<td>2.2233</td>
</tr>
<tr>
<td>ET</td>
<td>3 → 3, 3, → 3 → 3 → 0</td>
<td>0.30508</td>
<td>0.27232</td>
<td>1.1203</td>
</tr>
</tbody>
</table>

Results

All subsequences found under health concern discontinuations contained episodes of injectables (DHS Code 3) as the contraceptive method (Table 2). Additionally, patterns for DHS Kenya (2014), DHS Ghana (2014) and DHS Burkina Faso (2010) are discriminant (See Support Left, Support Right and Lift metrics) while for DHS Ethiopia (2016) the pattern containing injectables was not discriminant for the health concern discontinuations. Additional discriminatory subsequence experiments with other reasons for discontinuation can be seen in Table 3. We can observe that in both DHS Kenya (2014) and DHS Ghana (2014) discriminative patterns were found with consecutive use of Rhythm (DHS Code 8) as a contraceptive method in the group of calendars that reported to discontinue due to becoming pregnant while using a contraceptive method (DHS Discontinuation Code 1). For sequence mining visualization, a combination of directed graphs and bar plots (for support metrics) are used, with an example provided in Figure 3.
Table 3: Examples of discriminatory subsequences with particular discontinuation reasons. DHS Code Discontinuation Reasons (DR): Became pregnant while using a contraceptive method (DHS Discontinuation Code 1), wanted to become pregnant (DHS Discontinuation Code 2). Episodes: Rhythm (DHS Code 8), Pregnancy (DHS Code P).

<table>
<thead>
<tr>
<th>Discriminatory reason</th>
<th>Country</th>
<th>Subsequence</th>
<th>Support Left</th>
<th>Support Right</th>
<th>Lift</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) vs (2) KE</td>
<td>8 → P → P</td>
<td>0.26785</td>
<td>0.07596</td>
<td>3.5260</td>
<td></td>
</tr>
<tr>
<td>(1) vs (2) KE</td>
<td>8 → 8 → P</td>
<td>0.27142</td>
<td>0.07658</td>
<td>3.5440</td>
<td></td>
</tr>
<tr>
<td>(1) vs all other DR GH</td>
<td>8 → 8</td>
<td>0.30693</td>
<td>0.07666</td>
<td>4.0034</td>
<td></td>
</tr>
</tbody>
</table>

Figure 3: Discriminatory subsequences with support for Side effects/Health concerns (4) versus other discontinuation reasons (X) for DHS Ghana (2014). (A, B): The support value of the pattern for the left dataset and the right dataset respectively, are aggregated for all individual subsequences. i.e., the mean and standard deviation are computed. (C, D): Each episode e.g. Injectable, is repeated n number of times where n is indicated using the corresponding color intensity on the color bar (E). A subsequence may consist of a single repeated episode as in (D) or multiple episodes linked with an arrow as in (C).

Experimental setup for Q2: Contraceptive Transitions

Probability distributions of contraceptive transitions

To provide an overview of contraceptive use across each country based on the calendar data, the probability distributions per contraceptive type is estimated. The domain expert can define the time interval to use for the estimations, which can be a specific year of the calendar data or the totality of years available. To determine how long an individual used a particular contraceptive before switching to other methods or discontinuing to non-use, the frequencies of consecutive months of use for each specific contraceptive per country was calculated. Given the itemset of type of episodes, $D_i$ the set of sequences of calendar data for the country $i$, we extracted subsequences with one type of episode $j \in I - \{B, P, T, 0\}$. The length for each subsequence was then calculated, excluding episodes that do not correspond to a contraceptive method. As a result, a set of vectors is constructed $V_{j|i} = ([T_j], \cdots, [T_n])$, where $j \in I$ and $i \in D$ $|T_i|$ is the length of each subsequence found for that contraceptive over all calendar data in $D_i$. The Kernel Density Estimation (KDE) algorithm is used to estimate the probability density function applied to $\{V_{j|i} | j \in I - \{B, P, T, 0\}\}$. An example of interactive violin plots for consecutive months of contraceptive use distributions for DHS Kenya (2014) with six different contraceptive methods (five modern methods and one traditional method) is shown in Figure 2. This type of visualisation provides the domain expert with the tools to compare distributions across countries and methods. The probability of an individual using a particular method of contraception for a number of months is also estimated. Domain experts may find such analyses helpful for visualizing the switching patterns between different contraceptive;
not simply who discontinued or why, but also some insight on what the type of transition was (i.e., what type of contraception the person used next, if any). Furthermore, the probability distribution of consecutive episodes may allow experts to see and compare length of use in ways that show patterns not apparent from the usual “average months of use” that is commonly used.

After studying the trends for consecutive months of use for each country, a simple one-to-one pattern transition is provided. Using the subsequence already extracted in the previous section, ordered pairs such as a set of tuples $P_k = ((T_1, T_2), \ldots, (T_{n-1}, T_n))$ are formulated, where $T_i$ is a subsequence of one type of contraceptive $j$ with $j \in I$, $k \in D$ and $(T_1, T_2)$ is a transition pair from each subsequence found to other contraceptive methods in $D_k$. Examples of these transitions can be seen in Figure 4. In each Sankey plot, transitions from one type of episode, in this case an Injectable and Rhythm episode, to all the other types of episodes, may be observed. For women that transit, domain experts would find it useful to know what they transition to, in order to better understand contraceptive uptake. These one-step transitions are visualized using a Sankey plot, wherein the contraceptive of interest is selected as the ‘source’ of the flow and plot the amount of transition (indicated by the flow size) to all other ‘destinations’ i.e., contraceptive methods. The Sankey plots in Figure 4 illustrate the transitions from injectables in DHS Burkina Faso (2010). This analysis was repeated for all five countries and the results are summarized here for brevity.

**Results**

The percentages of transitions to non-use (no contraceptive used) constitute the largest proportion among all the countries studied: 85.36% in DHS Burkina Faso (2010), 57.39% in DHS Kenya (2014), 76.7% in DHS Ethiopia (2016), 70.42% in DHS Nigeria (2013) and 75.72% in DHS Ghana (2014). Importantly, injectables as a method of contraception had high rates of transition to episodes of non-use in all countries. In contrast, episodes of Rhythm use showed high rates of transition to both pregnancy and non-use in all countries. Figure 4 shows the transitions from injectables in DHS Burkina Faso (2010). The percentages to pregnancy episodes were: 57.38% for DHS Burkina Faso (2010), 59.67% for DHS Kenya (2014), 34.38% for DHS Ethiopia (2016), 55.36% for DHS Nigeria (2013) and 67.18% for DHS Ghana (2014). And the percentages to non-use were 34.43% for DHS Burkina Faso (2010), 16.57% for DHS Kenya (2014), 53.12% for DHS Ethiopia (2016), 30.36% for DHS Nigeria (2013) and 24.43% for DHS Ghana (2014). An interesting result that emerged was the bimodal distribution of Rhythm users compared to other methods, a finding that was not evident to the team until the episodes were visualized in this way (Figure 2). This finding implies that while some women briefly try the Rhythm method, a significant portion of them use this method for extended periods of time. Given that Rhythm has a high transition rate to unintended pregnancy, this insight draws attention to domain experts who aim to reduce the risk of unintended pregnancy.
Discussion

Our goal with this work was to illustrate the potential of sequence mining techniques to provide a data-driven approach to understanding contraceptive dynamics using the DHS survey data. This may inform policymakers and health workers as they address unmet needs in contraceptive usage. As such, this work provides domain experts with the ability to visualize and compare frequency distributions and discontinuation flows across different methods, sub-populations, and countries. Our analysis shows that patterns of continuous injectables episodes were a precursor to contraceptive discontinuation due to health concerns in four out of five datasets studied. These results need further examination to understand why these patterns appear and how the subpopulation that exhibits these sequences of episodes can be better characterized. Nonetheless, from the research literature, we know that side effects of modern family planning methods, either experienced or anticipated, have been identified as a common reason women choose not to start or discontinue contraceptives. Fear of side effects may occur when a woman or someone she knows has experienced side effects with a method when rumors are considered as facts, or rare complications are exaggerated\textsuperscript{20–23}. The temporal patterns found in this work complement existing research while providing a time-window that potential interventions can be made. For example, targeted education plans for subpopulations characterized by a pattern of contraceptive use identified result in discontinuation.

In another experimental setup, the domain experts aimed to create a “super dataset” by combining calendar data from multiple countries so as to perform a stratified DSM analysis such that there is sufficient sample size in rare classes. However, discriminatory patterns were hardly extracted from the combined datasets from different countries, which reflects the need to avoid blind merging of different datasets, with different years, countries and/or cultures. An approach of this kind will benefit from taking into account sub-populations that show some similarity in terms of their FP uptake or other factors. This has opened up a new research question: can we identify sub-populations that are “similar” enough to be able to combine their data, thus enabling sufficient data sizes for stratified analysis?

Conclusions

Our proposed approach provides subject-matter experts with insights based on data-driven automatic sequence mining and interactive visualization tools. This work analyzes a series of episodes in a contraceptive uptake pattern and identifying which sequences lead to discontinuation, to what extent these sequences are discriminatory, and whether these sequences are common across sub-populations. The proposed approach was tested across large datasets such as DHS Survey and Contraceptive Calendar Data. As more reliable calendar information becomes available in the future, such as PMA2020\textsuperscript{24}, this work demonstrates the potential to provide information to policymakers regarding contraceptive use and discontinuation for a given subpopulation. After finding contraceptive patterns on reliable data, subpopulations with the same patterns may be identified to perform targeted interventions. These interventions can range from educational material regarding a particular method to guarantee access to contraceptive methods after a given time-window.

Acknowledgements

We thank the Family Planning team from Bill & Melinda Gates Foundation, especially Jamaica Corker, Aparna Seth, Damian Walker and Uyi Stewart for their partnership with IBM Research Africa. We would also like to thank Kush Varshney, IBM T. J. Watson Research Center, Stephanie Müller and Sibusisiwe Makhanya, IBM Research Africa for their advice and feedback on this work. This work is funded by Bill & Melinda Gates Foundation, investment ID 52720.

References

3. Azuike E, Ikeako L, Ezeobi I, Ezebitalu I, Umeobika J, Obi K, et al. Predictors of discontinuation of contracep-
15. Jian Pei, Jiawei Han, Mortazavi-Asl B, Jianyong Wang, Pinto H, Qiming Chen, et al. Mining sequential patterns by pattern-growth: the PrefixSpan approach. IEEE Transactions on Knowledge and Data Engineering. 2004 Nov;16(11):1424–1440.
State of the Art Causal Inference in the Presence of Extraneous Covariates: A Simulation Study

Raluca Cobzaru\textsuperscript{1,2}, Sharon Jiang\textsuperscript{1,2}, Kenney Ng\textsuperscript{1,3}, Stan Finkelstein\textsuperscript{1,2}, Roy Welsch\textsuperscript{1,2}, Zach Shahn\textsuperscript{1,3}

\textsuperscript{1}MIT-IBM Watson AI Lab, Cambridge, MA, USA
\textsuperscript{2}Massachusetts Institute of Technology, Cambridge, MA, USA
\textsuperscript{3}IBM T. J. Watson Research Center, Yorktown Heights, NY, USA

Abstract

The central task of causal inference is to remove (via statistical adjustment) confounding bias that would be present in naive unadjusted comparisons of outcomes in different treatment groups. Statistical adjustment can roughly be broken down into two steps. In the first step, the researcher selects some set of variables to adjust for. In the second step, the researcher implements a causal inference algorithm to adjust for the selected variables and estimate the average treatment effect. In this paper, we use a simulation study to explore the operating characteristics and robustness of state-of-the-art methods for step two (statistical adjustment for selected variables) when step one (variable selection) is performed in a realistically sub-optimal manner. More specifically, we study the robustness of a cross-fit machine learning based causal effect estimator to the presence of extraneous variables in the adjustment set. The take-away for practitioners is that there is value to, if possible, identifying a small sufficient adjustment set using subject matter knowledge even when using machine learning methods for adjustment.

Introduction

In the absence of randomized trials, it is common for researchers to attempt to estimate treatment effects using observational healthcare data (e.g. from an insurance claims or electronic health record – EHR – database). Under strong assumptions, statistically adjusted comparisons of outcomes in patients receiving the treatment alternatives of interest provide unbiased estimates of the results that would be observed in the hypothetical randomized trial researchers wish they could run.

The central aim of statistical adjustment is to remove confounding bias that would be present in naive unadjusted comparisons. Statistical adjustment can be broken down into two steps. In the first step, the researcher selects some set of variables to adjust for. In the second step, the researcher implements a causal inference algorithm to adjust for the selected variables and estimate the average treatment effect (ATE). In this paper, we use a simulation study to explore the operating characteristics and robustness of state-of-the-art methods for step two (statistical adjustment for selected variables) when step one (variable selection) is performed in a realistically suboptimal manner. (Simulation studies to address questions like this are common and crucial in causal inference\textsuperscript{1–5} in particular, since counterfactual ground truth under alternative treatment strategies is rarely available in real data.)

If the selected set of variables to adjust for in an observational study is appropriate for removing bias, then the exchangeability assumption is said to hold\textsuperscript{[2]}. Exchangeability is partially an assumption about the richness of the database in that the assumption cannot hold if the database does not contain sufficient confounding variables. Exchangeability is also partially an assumption about the researcher’s ability to select the right variables to adjust for out of those available. Even if a database contains a sufficient adjustment set, a researcher may fail to adjust for a variable in the set or adjust for an extraneous variable that actually induces bias (so-called ‘M-bias’\textsuperscript{4,7}, which we will discuss in more detail later). Because it is not empirically verifiable from data whether exchangeability holds for a given set of variables, and because it is possible to run afoul of exchangeability by choosing either too many or too few variables to adjust for, variable selection is a very challenging problem\textsuperscript{3}. In practice, researchers typically err on the side of adjusting for too many variables and risking M-bias, rather than adjusting for too few and risking bias due to unobserved confounding. This approach has the effect that even if a selected adjustment set satisfies exchangeability and does not induce M-bias, it will often contain extraneous variables not strictly necessary to adjust for to estimate the treatment effect. Such extraneous variables are theoretically inconsequential from a bias perspective. But in practice, with finite data, extraneous variables of any kind may influence performance.

334
Once an adjustment set is selected, most statistical adjustment procedures require fitting so-called ‘nuisance’ models, i.e. regression models predicting treatment conditional on the covariates and regression models predicting outcome conditional on treatment and the covariates. These nuisance models are not directly of interest (hence the name), but rather they serve as inputs to treatment effect estimators. ‘Standardization’ or ‘g-computation’ estimators depend only on a nuisance model for the outcome, which must be correctly specified. Inverse probability of treatment weighted estimators depend only on a nuisance model for the treatment, which must be correctly specified. Doubly robust estimators take as inputs nuisance models for both the outcome and the treatment, but the estimators are consistent if at least one of the two nuisance models is correctly specified. Realistically, when data generating processes are nonlinear or complex, parametric nuisance models (such as generalized linear models) are bound to be misspecified. Flexible nonparametric data adaptive machine learning nuisance regression models have the desirable property that they will not be misspecified and will lead to unbiased treatment effect estimates asymptotically. But machine learning models converge slowly to the truth, leading to poor confidence interval coverage and significant bias when they are used to estimate nuisance models in traditional effect estimators with datasets of realistic size. Recently, though, it was shown that doubly robust estimators computed via cross-fitting and using machine learning outcome and treatment nuisance models behave as well (i.e. converge to the truth at the same rate) as estimators using correctly specified parametric models. (Cross-fitting means splitting the data into two portions, generating two estimates by plugging nuisance model estimates from one half of the data into effect estimators trained on the other half, and then averaging the estimates.) This approach constitutes the state-of-the-art, as it greatly ameliorates misspecification bias without sacrificing fast convergence rates.

A recent simulation study nicely illustrated the utility of cross-fitting with machine learning in a realistic setting. The authors generated data for a (simulated) observational study assessing the effects of statins on atherosclerotic cardiovascular disease (ASCVD). Their data generating process included four confounding variables based on treatment guidelines for the management of blood cholesterol. These confounders were generated to be interpreted as age, cardiovascular risk score, diabetes, and blood cholesterol levels. Probability of statin use was generated as a complex nonlinear function of the confounders, and probability of ASCVD incidence was generated as a complex nonlinear function of the confounders and statin use. Treatment effect estimators based on simple parametric nuisance models that were (inevitably) misspecified were biased for the true effect. Effect estimators based on flexible nonparametric machine learning nuisance models (but not doubly robust and estimated via cross-fitting) were also biased. Only doubly robust cross fit machine learning estimators were found to be unbiased in this realistic but relatively simple setting. While this simulation demonstrated the necessity of machine learning with cross-fitting, it did not explore the approach’s robustness to settings where the small set of truly important confounders is not known a priori.

In this paper, we build on the above study by evaluating performance of cross-fitting with machine learning in the (common in practice) presence of extraneous covariates. We find that performance can degrade as extraneous variables are added, with the degree of degradation depending on properties of the extraneous variables. We considered extraneous variables that are pure noise, weakly or strongly associated with the outcome but not the exposure, weakly or strongly associated with the exposure but not the outcome (except possibly through the exposure), or weakly or strongly associated with both outcome and exposure so as to induce M-bias. Generally, this simulation study highlights potential drawbacks of the tempting adjustment strategy: “throw everything in and let the machine learning models sort it out”. There is real value to, if possible, identifying a small sufficient adjustment set using subject matter knowledge, even when using state-of-the-art machine learning methods for adjustment.

Methods

The data generating process developed in Zivich and Breskin provides the core of all of our simulations. In each simulation, we first generate the variables $Z$, $X$, and $Y$ as in their study. $Y$ denotes the observed binary outcome variable, to be interpreted as incidence of atherosclerotic disease. $X$ denotes the treatment, to be interpreted as statin use. $Z$ is a vector of four confounding variables meant to be interpreted as age, cardiovascular risk, diabetes, and cholesterol levels. We denote the individual confounders as $Z_{age}$, $Z_{risk}$, $Z_{diab}$, and $Z_{chol}$. We first generate $Z$ as
described in the appendix of Zivich and Breskin. We next generate $X$ given $Z$ according to the treatment model:

$$
\Pr(X = 1|Z) = \text{Bernoulli}(\exp( -3.471 + 1.390 Z_{\text{diab}} + 0.112 Z_{\text{chol}} + 0.973 I(Z_{\text{chol}} > \log(60))
- 0.046(Z_{\text{age}} - 30) + 0.003(Z_{\text{age}} - 30)^2 + 0.273 I(0.05 \leq Z_{\text{risk}} < 0.075)
+ 1.592 I(0.075 \leq Z_{\text{risk}} < 0.2) + 2.46 I(Z_{\text{risk}} \geq 0.2)))
$$

This treatment model was designed based on the primary prevention guidelines for management of blood cholesterol such that statin prevalence matched empirical results in US adults. Let $Y(x)$ denote the counterfactual outcome under treatment assignment $X = x$. We generate $Y(x)$ given $Z$ from the outcome model:

$$
\Pr(Y(x) = 1|Z) = \text{Bernoulli}(\exp(-6.250 - 0.75x + 0.35x(5 - Z_{\text{chol}})I(Z_{\text{chol}} < \log(130)) + 0.45(Z_{\text{age}} - 39)^{0.5}
+ 1.75Z_{\text{diab}} + 0.29 \exp(Z_{\text{risk}} + 1) + 0.14 Z_{\text{chol}}^2 I(Z_{\text{chol}} > \log(120)))
$$

and then set the observed outcome $Y$ to be the counterfactual outcome corresponding to the received treatment, i.e. $Y = XY(1) + (1 - X)Y(0)$.

We are interested in estimating the average treatment effect $E[Y(1)] - E[Y(0)]$ from the observed data $Z, X, Y$. The data generating process satisfies the three causal assumptions necessary to do so:

1. Consistency: $Y = Y(X)$
2. Exchangeability: $Y(x) \perp X|Z$ for all $x$
3. Positivity: $\Pr(X = x|Z = z) > 0$ for all $x, z$ such that $\Pr(Z = z) > 0$

However, in addition to these causal assumptions holding, for treatment effect estimators to be unbiased it is also necessary to accurately specify the nuisance treatment and/or outcome models (1) and (2). Because these models are (realistically) complex, standard parametric model specifications would inevitably be incorrect. Therefore, it is necessary to use nonparametric machine learning methods to estimate the treatment and outcome models. (Zivich and Breskin showed that parametric specifications of the nuisance models using just the main effect terms for the four confounders led to biased estimates.) And when plugging in machine learning estimates of nuisance models to a treatment effect estimator with cross-fitting to attain unbiased estimates and nominal coverage under minimal assumptions. Indeed, in this particular data generating process, machine learning with cross-fitting was the only approach that yielded unbiased estimate.

The focus of our simulation study is on the robustness of cross-fit machine learning estimators to the presence of a vector of extraneous covariates $L$. We roughly follow Brookhart et al. and Liu et al. in considering extraneous variables of varying types – pure noise independent of both treatment and outcome, associated with treatment but not outcome (except possibly through the treatment), associated with the outcome but not the treatment, and associated with both outcome and treatment so as to induce M-bias. Graphical models describing the dependence structure for each extraneous variable type are shown in Figure 1.

We also consider varying strength of association and number and mix of extraneous variables, as illustrated in the schematic in Figure 2. Table containing our results, enumerates the simulation settings we evaluated. Below, we describe in detail how we generated the extraneous variables for each simulation setting.

1) Noise variables ($n = 10, 20$): Continuous variables were generated as independent identically distributed $\mathcal{N}(0, 1)$ and binary variables were generated as independent identically distributed Bernoulli(.5). As illustrated in panel (A) of Figure 1, these variables are completely independent of treatment ($X$), confounders ($Z$), and outcome ($Y$). Assumptions 1-3 required for causal inference would still hold in this dataset with $\{Z, L\}$ as the adjustment set.
2) Variables associated with exposure $X$ but not the outcome $Y$, except through exposure $(n = 10, 20; \text{association} = \text{‘strong’, ‘weak’})$: First, we generate a binary intermediary variable $U$ associated with exposure as $U \sim \text{Bernoulli}(\expit(\beta X))$. Using the same $U$, we generate each binary component of $L$ as $\text{Bernoulli}(\expit(\beta U))$. Using the same $U$, we also generate each continuous component of $L$ as $\mathcal{N}(U, \sigma^2)$. The parameters used are

$$
\beta = \begin{cases} 
0.65 & \text{if weak association} \\
1.55 & \text{if strong association} 
\end{cases} \quad \text{and} \quad \sigma = \begin{cases} 
2 & \text{if weak association} \\
0.5 & \text{if strong association} 
\end{cases}
$$

See Figure 1 panel (B) for a graph depicting the dependencies in this data generating process. We chose $\beta$ such that the coefficient of a logistic regression of $X$ on $L$ would be equal to 0.5 in the strong case and 0.1 in the weak case. These coefficients approximately matched the 80th and 20th percentiles, respectively, of the distribution of log odds ratios of baseline variables (indicators of all pre-treatment diagnoses and medications) and exposures in several trial emulation cohorts we constructed from a claims database for another study. Assumptions 1-3 required for causal inference would still hold in this dataset with $\{Z, L\}$ as the adjustment set in place of $Z$.

Note that generating $L$ as a function of $U$ has several consequences. First, it induces dependence among the various components of $L$. Second, it caps the strength of the joint association between all elements of $L$ and $X$ by the strength of association between $U$ and $X$. Observing 20 independent variables with strong associations with exposure would allow us to almost perfectly predict exposure. But observing 20 dependent variables (linked by $U$), each with a strong marginal association with exposure, only allows us to predict exposure at best as well as we could by directly observing $U$. This is a desirable property in our simulation as it is not usually possible to perfectly predict exposure in healthcare datasets.

3) Variables associated with outcome $Y$ but not exposure $(n = 10, 20; \text{association} = \text{‘strong’, ‘weak’})$: First, we generate a binary intermediary variable $U$ associated with the counterfactual untreated outcome as $U \sim \text{Bernoulli}(\expit(\beta Y(0)))$. Using the same $U$, generate each binary component of $L$ as $\text{Bernoulli}(\expit(\beta U))$. Using the same $U$, we additionally generate each continuous component of $L$ as $\mathcal{N}(U, \sigma^2)$. The parameters used are

$$
\beta = \begin{cases} 
0.65 & \text{if weak association} \\
1.55 & \text{if strong association} 
\end{cases} \quad \text{and} \quad \sigma = \begin{cases} 
2 & \text{if weak association} \\
0.5 & \text{if strong association} 
\end{cases}
$$

Assumptions 1-3 required for causal inference would still hold in this dataset with $\{Z, L\}$ as the adjustment set in place of $Z$. See Figure 1 panel (C) for a graph depicting the dependencies in this data generating process. As for the exposure associated variables, we chose $\beta$ such that the coefficient of a logistic regression of $Y$ on $L$ would be equal to 0.5 in the strong case and 0.1 in the weak case.

As in the exposure associated variable data generating process, including the intermediate variable $U$ induces dependence among components of $L$ and prevents perfect prediction of the (counterfactual) outcome even in the presence of many variables that are each marginally strongly associated with the outcome.

4) Variables associated with treatment $X$ and outcome $Y$ so as to induce M-bias $(n = 10, 20; \text{associations} = \text{‘strong’, ‘weak’})$: We generate hidden exposure-associated variable $U_1$ as $\text{Bernoulli}(\expit(\beta X))$ and hidden outcome-associated variable $U_2$ as $\text{Bernoulli}(\expit(\beta Y(0)))$. Then, we generate each continuous component of $L$ as $\mathcal{N}(U_1 + U_2, \sigma^2)$ and each binary component of $L$ as $\text{Bernoulli}(\expit(\beta (U_1 + U_2)))$, where

$$
\beta = \begin{cases} 
0.7 & \text{if weak association} \\
1.65 & \text{if strong association} 
\end{cases} \quad \text{and} \quad \sigma = \begin{cases} 
1.75 & \text{if weak association} \\
0.8 & \text{if strong association} 
\end{cases}
$$

See Figure 1 panel (D) for a graph depicting the dependencies in this data generating process. As for exposure and outcome associated variables, we chose $\beta$ such that the coefficients of logistic regressions of $X$ and $Y$ on each component of $L$ would be equal to 0.5 in the strong case and 0.1 in the weak case.

Under the M-bias data generating process, the exchangeability assumption is not satisfied with $\{Z, L\}$ as the adjustment set in place of $Z$. Adjusting for $L$ induces so-called \’collider bias\’ generating a spurious association between $X$ and $Y$. Thus, bias observed under this setting is not just due to degradation of the estimator’s
Table 1: Descriptive statistics of a single dataset generated under the mixed setting (5) with \( n = 20 \) extra covariates

<table>
<thead>
<tr>
<th></th>
<th>( X = 1 ) (( n = 1178 ))</th>
<th>( X = 0 ) (( n = 1822 ))</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>% or mean (SD)</td>
<td>% or mean (SD)</td>
</tr>
<tr>
<td>( Z_{\text{age}} )</td>
<td>55 (9.6)</td>
<td>52 (7.3)</td>
</tr>
<tr>
<td>( Z_{\text{chol}} )</td>
<td>4.89 (0.18)</td>
<td>4.85 (0.18)</td>
</tr>
<tr>
<td>( Z_{\text{risk}} )</td>
<td>0.14 (0.19)</td>
<td>0.06 (0.06)</td>
</tr>
<tr>
<td>( Z_{\text{diab}} )</td>
<td>41%</td>
<td>7%</td>
</tr>
<tr>
<td>( Z_{\text{noise}} )</td>
<td>0.008 (0.98)</td>
<td>0.003 (1.01)</td>
</tr>
<tr>
<td>( Z_{\text{exposure}} )</td>
<td>0.66 (2.04)</td>
<td>0.44 (2.00)</td>
</tr>
<tr>
<td>( Z_{\text{exposure}} )</td>
<td>0.79 (0.64)</td>
<td>0.48 (0.72)</td>
</tr>
<tr>
<td>( Z_{\text{outcome}} )</td>
<td>0.61 (2.05)</td>
<td>0.54 (2.10)</td>
</tr>
<tr>
<td>( Z_{\text{outcome}} )</td>
<td>0.64 (0.71)</td>
<td>0.59 (0.69)</td>
</tr>
<tr>
<td>( Z_{\text{M-bias}} )</td>
<td>1.20 (1.81)</td>
<td>1.07 (1.88)</td>
</tr>
<tr>
<td>( Z_{\text{M-bias}} )</td>
<td>1.50 (1.01)</td>
<td>1.11 (1.05)</td>
</tr>
<tr>
<td>( Y )</td>
<td>31%</td>
<td>27%</td>
</tr>
</tbody>
</table>

We produced 2,000 datasets with 3,000 observations containing extraneous variables of each type listed above. Table 1 illustrates the characteristics of a single simulation sample under the mixed setting with \( n = 20 \) extra covariates, with an example of each extraneous variable type. In each dataset, we estimated the average treatment effect using the doubly robust augmented inverse probability weighted estimator via cross fitting. We used ensembles of regression models (multi-layer perceptron with a hidden layer containing 4 nodes, a random forest with 500 trees, logistic regression without regularization, empirical mean, generalized additive model with 4 splines, and a generalized additive model with 6 splines) to estimate the nuisance treatment and outcome models. Results for each simulation setting are summaries of performance of the estimator over the 2,000 datasets generated under that setting. All simulation code for our analyses can be found at https://github.mit.edu/rcobzaru/simulation-study.

Results

Results are summarized in Table 2. As in Zivich and Breskin, the cross-fit machine learning estimator produces unbiased estimates of the treatment effect when only true confounding variables were adjusted for. Adding extraneous noise variables led to moderate degradation in performance. Adding variables weakly associated with the treatment but not the outcome did not add more bias than noise variables. Adding variables strongly associated with the treatment but not the outcome did significantly increase bias compared to adding noise variables. Adding variables weakly associated with the outcome but not the treatment led to approximately the same bias increase as noise variables. Adding variables strongly associated with the outcome but not the treatment led to substantially less additional bias than noise variables, with 20 strongly outcome-associated variables leading to no additional bias at all. Adding variables with only weak M-bias inducing associations did not cause more bias than just adding noise variables. Predictably, adding variables with strong M-bias inducing associations led to the largest increase in bias of the estimator. A mix of extraneous variables of all types also induced bias in line with what might be expected from the individual mixture components. Standard errors and corresponding confidence interval widths were increased in the presence of strong variables.
Figure 1: Directed Acyclic Graphs describing the relationships of the four types of extraneous covariates ($L$) to treatment ($X$), outcome ($Y$), and confounders ($Z$).

Figure 2: Schematic of the space of simulation studies we explore
Table 2: Each row corresponds to one simulation setting described in the Methods section. The first three columns jointly define the simulation setting. Bias is the difference between the mean effect estimate over 2,000 datasets and the true effect (as computed via a simulation of a dataset with 1,000,000 observations). Absolute bias as percentage of true effect is just $100 \times \text{bias}/(\text{true effect size})$. RMSE is the root mean squared error of the 2,000 effect estimates. ESE is the empirical standard error of the 2,000 estimates. CLD is the average ‘confidence limit difference’ of the confidence intervals over 2,000 datasets. Coverage is the proportion of the 2,000 confidence intervals that contained the true effect size.

<table>
<thead>
<tr>
<th>Extra covariate types</th>
<th>Number of extra covariates</th>
<th>Strength of association</th>
<th>Bias</th>
<th>Absolute bias as % of true effect</th>
<th>RMSE</th>
<th>ESE</th>
<th>CLD</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>0</td>
<td>None</td>
<td>-0.001</td>
<td>1</td>
<td>0.017</td>
<td>0.017</td>
<td>0.070</td>
<td>96.2%</td>
</tr>
<tr>
<td>Noise</td>
<td>10</td>
<td>None</td>
<td>-0.007</td>
<td>6</td>
<td>0.019</td>
<td>0.018</td>
<td>0.071</td>
<td>93.6%</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>None</td>
<td>-0.008</td>
<td>7</td>
<td>0.023</td>
<td>0.021</td>
<td>0.071</td>
<td>91.1%</td>
</tr>
<tr>
<td>Exposure</td>
<td>10</td>
<td>Weak</td>
<td>-0.006</td>
<td>5</td>
<td>0.019</td>
<td>0.018</td>
<td>0.071</td>
<td>94.8%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Strong</td>
<td>-0.010</td>
<td>9</td>
<td>0.021</td>
<td>0.019</td>
<td>0.076</td>
<td>91.8%</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>Weak</td>
<td>-0.007</td>
<td>7</td>
<td>0.019</td>
<td>0.018</td>
<td>0.071</td>
<td>92.6%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Strong</td>
<td>-0.014</td>
<td>13</td>
<td>0.024</td>
<td>0.019</td>
<td>0.076</td>
<td>89.5%</td>
</tr>
<tr>
<td>Outcome</td>
<td>10</td>
<td>Weak</td>
<td>-0.005</td>
<td>5</td>
<td>0.019</td>
<td>0.018</td>
<td>0.071</td>
<td>94.2%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Strong</td>
<td>-0.003</td>
<td>3</td>
<td>0.018</td>
<td>0.018</td>
<td>0.070</td>
<td>94.4%</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>Weak</td>
<td>-0.007</td>
<td>7</td>
<td>0.019</td>
<td>0.017</td>
<td>0.071</td>
<td>93.7%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Strong</td>
<td>-0.001</td>
<td>1</td>
<td>0.018</td>
<td>0.018</td>
<td>0.070</td>
<td>93.8%</td>
</tr>
<tr>
<td>M-bias</td>
<td>10</td>
<td>Weak</td>
<td>-0.008</td>
<td>8</td>
<td>0.020</td>
<td>0.018</td>
<td>0.071</td>
<td>93.0%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Strong</td>
<td>-0.019</td>
<td>17</td>
<td>0.025</td>
<td>0.019</td>
<td>0.071</td>
<td>80.9%</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>Weak</td>
<td>-0.013</td>
<td>12</td>
<td>0.021</td>
<td>0.017</td>
<td>0.071</td>
<td>91.7%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Strong</td>
<td>-0.022</td>
<td>20</td>
<td>0.028</td>
<td>0.018</td>
<td>0.071</td>
<td>80.2%</td>
</tr>
<tr>
<td>Mix</td>
<td>10</td>
<td>Mixed</td>
<td>-0.014</td>
<td>12</td>
<td>0.023</td>
<td>0.018</td>
<td>0.073</td>
<td>88.1%</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>Mixed</td>
<td>-0.018</td>
<td>17</td>
<td>0.027</td>
<td>0.019</td>
<td>0.074</td>
<td>81.5%</td>
</tr>
</tbody>
</table>

exposure-associated variables, but otherwise similar across settings. Thus, degradation of confidence interval coverage below the nominal 95% rate stemmed mainly from bias.

Conclusion

While optimal variable selection for causal inference is theoretically possible with detailed knowledge of the causal graph underlying the data generating process\cite{16} in practice it is an intractable problem\cite{8}. It is not empirically verifiable from data whether a given adjustment set satisfies the exchangeability assumption, and a variable that would induce M-bias is statistically indistinguishable from a confounder that is important to adjust for. Thus, without excellent subject matter knowledge, adjustment sets are bound to be imperfect. If a researcher is able to identify a set of variables they are confident contains a sufficient adjustment set, it probably contains extraneous variables as well. Past simulation studies have produced insights and heuristics about the consequences of extraneous variables for causal inference. However, it is not known to what degree these general heuristics extend to settings in which recently introduced cross-fit doubly robust estimators with machine learning nuisance models are required. How well can cross-fit machine learning effect estimators discard (or utilize) extraneous adjustment variables?

Our results were largely consistent with past simulation studies. In particular, Brookhart et al\cite{2} found that adjusting for variables associated with exposure but not the outcome could increase variance and inflate unmeasured confounding bias. This result was confirmed and generalized theoretically by Ding et al\cite{17}. In our extraneous exposure-associated variable simulation setting, there was no structural bias due to unmeasured confounding to inflate, i.e. the exchangeability assumption held. Presumably, if we had generated very large data sets under this setting the bias would have shrunk to zero. However, in our realistically sized data sets, there was some residual confounding bias due to failure to properly statistically adjust for the observed confounding variables, especially in the presence of extra adjustment variables. This residual confounding bias was inflated by adjustment for extraneous strongly exposure-associated...
variables.

Past simulation studies have further shown that adding variables associated with the outcome is generally beneficial, reducing variance without inducing bias. Our results were somewhat consistent with this finding, with extraneous strongly outcome-associated variables inducing little to no bias. However, weakly outcome-associated variables did induce some bias, possibly reflecting that when it is difficult to properly adjust for the actual confounders due to complex nuisance models, ‘distractions’ in the form of extraneous variables of any kind can be detrimental.

Of course, it is not surprising that M-bias inducing variables lived up to their name. There was no expectation that machine learning could magically compensate for violation of the exchangeability assumption required for causal inference. It might be reassuring that, at least in this setting (designed to be fairly realistic), M-bias from 10 variables with weak associations with exposure and outcome was hardly stronger than the bias induced by extraneous noise variables.

We might have hoped that pure noise variables would have been discarded by the machine learning algorithms and thus caused undetectably little bias. Unfortunately, it seems that even adding a relatively small number of noise variables ‘distracts’ the nuisance models enough to induce moderate residual confounding bias with a sample size of 3,000.

We believe our ‘Mixed’ simulation setting offers a roughly realistic picture of the impact of extraneous variables one might see in practice. Bias is non-negligible, enough to highlight the potential importance of strong subject matter knowledge to winnow down the adjustment set but without causing too much pessimism at the prospect of reliable causal inference from observational data.

Of course, in all of our simulation settings the true set of confounders was contained in the adjustment set, which would seldom be the case in a real study. But these results still suggest some guidance for variable selection (and, from an informatics perspective, collection) with state-of-the-art machine learning causal effect estimators if nuisance models are thought to be complex. In particular, the simulations illustrate that when nuisance models are complex the (common) strategy of ‘throwing everything in’ for the machine learning models to sort out comes with a price. Extraneous variables provide little to no upside in this setting (unlike the simple nuisance model case, where outcome-associated variables are net beneficial) and potentially significant downside (exposure-associated variables induce bias even in the absence of unmeasured confounding, M-bias lurks). The simulation does not offer guidance on how to trade off the risks of extraneous variables against the risks of omitted variable bias in a given application, but makes clear that the tradeoff does still exist even when using machine learning for adjustment. This is contrary to the attitude (that we have often heard expressed and even held ourselves) that, apart from M-bias, including extraneous variables is inconsequential when using machine learning to fit nuisance models. The primary contribution of machine learning to causal inference should probably be viewed as fitting flexible nuisance models to adjust for carefully selected covariates, not variable selection. That said, apart from rather extreme cases of M-bias, effect estimates were pretty reasonable across settings, and this study is not cause for despair for researchers who lack strong subject matter knowledge to guide variable selection and nonetheless find themselves in the fortunate situation of believing that available covariates contain a sufficient adjustment set.

References


Vaping at the VA: Developing an Annotated Corpus of Electronic Cigarette Mentions in Clinical Notes at the Department of Veterans Affairs

Mike Conway, MSc, PhD1, Patrick R Alba, MS2,3, Shu-Hong Zhu, PhD4, Olga V Patterson, PhD2,3

1 Department of Biomedical Informatics, University of Utah, Salt Lake City, UT; 2 Department of Internal Medicine, University of Utah, Salt Lake City, UT; 3 VA Salt Lake City Health Care System, Salt Lake City, UT; 4 Herbert Wertheim School of Public Health & Human Longevity Science, University of California San Diego, La Jolla, CA

Abstract

Use of Electronic Nicotine Delivery Systems (ENDS, colloquially known as “electronic cigarettes”) has increased substantially in the United States in the decade since 2010. However, currently relatively little is known regarding the documentation of ENDS use in clinical notes. With this study, we describe the development of an annotation scheme (and associated annotated corpus) consisting of 4,351 ENDS mentions derived from Department of Veterans Affairs clinical notes during the period 2010-2020. Analysis of our corpus provides important insights into ENDS documentation practices at the VA, in addition to providing a resource for the future development and validation of Natural Language Processing algorithms capable of reliably identifying ENDS-use status.

Introduction & Motivation

Electronic cigarettes — e-cigarettes, e-cigs, vapes, or Electronic Nicotine Delivery Systems (ENDS) — have exploded in popularity over the last ten years in the United States. It is estimated that by 2012, 75% of American adults had heard of ENDS, and 8.1% had tried them.1 By 2014, 94% of Americans had some awareness of ENDS, and 12.6% had tried them.2,3 Despite the growth in popularity of the product, little consensus currently exists regarding the safety of ENDS devices, with regulatory authorities, professional associations, and individual clinicians divided as to whether ENDS constitute a valuable smoking cessation tool,4,5 or rather are a potentially harmful technology that risks eroding hard-won achievements in denormalizing smoking.6,7 This uncertainty regarding the potential safety risk of ENDS use was further amplified by the 2019 outbreak of E-cigarette or Vaping Product Use-Associated Lung Injury (EVALI) in the United States,8 a development that sparked calls for enhanced regulation and surveillance of the products.9

Clinical practices regarding the documentation of ENDS use in the Electronic Health Record (EHR) are currently poorly understood, and what evidence that does exist regarding documentation patterns suggests that ENDS use is massively under-documented.10,11 This under-documentation is perhaps partially due to the absence of a standardized means of recording ENDS use in EHR systems,12 a situation that — at least for some health systems — has recently seen changes with the introduction of specific fields for capturing ENDS use.13 However, it holds true that, in the majority of cases, ENDS documentation, if it exists at all, is dispersed across clinical notes, thus necessitating the application of Natural Language Processing (NLP) methods or manual abstraction to extract relevant information. Further, the type of documentation commonly found in clinical notes (e.g. type of ENDS device, duration of ENDS use, frequency of ENDS use, ENDS start date, ENDS end date) is currently not well understood.

With this study, we describe the development of an annotation scheme (and associated annotated corpus) consisting of 4,351 ENDS-related mentions derived from Department of Veterans Affairs (VA) clinical notes during the period 2010 to 2020, in order to better understand how clinicians document ENDS use, and to develop a validated resource for the future training and evaluation of NLP algorithms to execute the task of automatically determining ENDS use status.

Background

The VA EHR provides a unique national dataset with which to investigate ENDS documentation and usage patterns, given that it is the largest healthcare system in the United States with 1,255 healthcare facilities nationwide and is used by over nine million veterans per year.12 The VA patient demographic is particularly interesting from a tobacco
The VA collects data regarding patient smoking history and current smoking status using several different approaches at the time of patient encounter. Chief among these is the utilization of Health Factors. However, health factors cannot easily be used to reliably identify the type of tobacco product used, and — most salient to our research question — whether a patient is an ENDS user. In the VA context, this ENDS-use information is typically found embedded in clinical text.

Automatically identifying smoking status using the EHR is a well-developed research area that has utilized a number of different methods, including the use of structured data, semi-structured data, and NLP approaches. However, there has been relatively little work focused on how ENDS use — as opposed to combustible tobacco use — is currently documented in the EHR. Notable exceptions include ENDS annotation scheme and corpus development work at Fairview Health System in Minneapolis and analyses of ENDS documentation using keyword matching methods conducted at the VA and on a larger scale, at Kaiser Permanente. All these health systems exhibited an increase in ENDS use documentation (i.e. the prevalence of ENDS-related keywords) over time, but notable, this increase is below what would be expected given the prevalence of ENDS use in the general population, indicating that ENDS use is systematically under-documented in these systems.

Our overarching aim with this study is to analyze ENDS documentation patterns in VA clinical notes, including how ENDS documentation practices have changed over time. More specifically, we report on the development of an annotation scheme (and a manually annotated corpus based on that annotation scheme) suitable for the annotation of ENDS mentions in VA clinical notes. Our resulting corpus consisted of 4,351 annotations across sixteen annotation categories related to ENDS usage (e.g. potential-user, passive-exposure, attempting-to-quit-ENDS) that can be utilized as a resource for the future development and validation of NLP algorithms. The contribution of this work consists in (a) the development of validated annotation guidelines for annotating ENDS mentions in VA clinical notes and (b) based on this annotation scheme, the development of a high-quality manually annotated corpus that provides important insights into the range of current and historical ENDS use documentation practices, in addition to providing a resource for the future development and validation of NLP algorithms. The development of an annotated corpus is a necessary condition for the creation and validation of an NLP algorithm capable of reliably determining ENDS use status, which in turn is required to interrogate EHR systems regarding key open epidemiological questions concerning ENDS use (e.g. what proportion of ENDS users quit tobacco? are clinicians recommending ENDS use as a cessation method?)

Materials & Methods

Our initial cohort consisted of VA patients with a history of smoking (as determined by VA Health Factors) who had utilized clinical services between 2010 and 2020. To better understand how ENDS terminology has changed over time, we used a set of ENDS-related keywords derived from various sources, including previous work, additional iterative corpus data analysis, and a word-embedding model trained on the totality of VA notes. This process resulted in 77 keywords that were categorized into five broad groups (electronic cigarette, e-cigarette, e-cig, vape) or other keywords (other variants). Using these keywords, we performed a search across our corpus of 4,604,856 million patients in order to allow us to identify the prevalence of ENDS keywords and observe changes in ENDS terminology over time. This process resulted in the identification of 1,638,884 ENDS mentions for 418,170 patients across 1,191,133 clinical notes. From our cohort of VA patients, we randomly sampled notes containing ENDS mentions. We then iteratively developed an annotation scheme for VA ENDS use documentation, using prior work reported in Winden et al. as our starting point. Four individuals participated in the annotation scheme construction process (authors MC & PA, in addition to two annotators with a background in nursing). The resulting annotation scheme consisted of five top level annotation categories (Active-User, Usage-Unknown, Irrelevant, Former-User, and Non-User) and sixteen lower level categories (see Figure for a graphical representation of the annotation scheme workflow, with additional examples provided in Table). In addition to annotation type, each ENDS mention is associated with demographic characteristics derived from structured data in the VA EHR (including gender, urban/rural location, outpatient/inpatient status, and age range). Further, each annotation is associated with its source note type (e.g. primary care, mental health,
Figure 1: Annotation scheme flowchart
As can be seen from Figure 2, there have been striking changes in ENDS-related terminology since 2010, with “vape”-related keywords superseding previously dominant ENDS keywords like “e-cigarette”.‡ Given this observation — i.e. that keywords used to refer to ENDS in clinical documentation shifted substantially over time — we elected to construct a stratified sample of ENDS mentions for the development of our annotated corpus. The initial stratified set of documents was created by selecting 200 random ENDS mentions every year from 2010-2020 (n = 2,200). In order to account for the increased variation and changes in usage in recent years, a second set of instances was selected using 250 randomly selected mentions every year from 2016-2020 (n = 1,250). The third and final set of annotations was selected to be held out for final testing of any NLP algorithms. This third set was selected using a random set of ENDS mentions proportionate to the total number of mentions identified each year (n = 901). In total, our annotated corpus consists of 4,351 ENDS mentions. Annotations were completed at the mention-level using Chex,27,28 a web-based, VA-developed application. Chex provides a simple user interface that allows annotators to view a small snippet of text surrounding an anchored keyword with the ability to expand to a complete document for further context when needed.

In order to ensure annotation quality, we double annotated sections of the corpus (10% of the training set and 100% of the test set), achieving comparable agreement scores to those gained during the annotation scheme development process. We computed agreement using Cohen’s kappa.29 Our annotation effort resulted in 4,351 ENDS mentions, derived from 4,345 clinical notes and 4,213 individual patients between 2010 and 2020.§

Results

Using our cohort of 4,604,856 patients we identified 1,638,884 ENDS mentions for 418,170 patients across 1,191,133 clinical notes in the VA EHR and observed rapid growth in both the volume and variety of ENDS-related terms over time (see Figure 2). The total volume of clinical notes containing ENDS-related keywords exhibited an average annual growth rate of 52.8% over the 10 year period, with 8,234 clinical notes identified in 2010 and 202,842 in 2019; this is

‡Note that a simple string matching approach was used to determine prevalence of a predetermined list of ENDS-related keywords that occurred in notes over the period 2010-2020. See previous footnote for a link to the complete set of keywords used.

§In line with our approved University of Utah Institutional Review Board protocol, and VA policy, the annotated corpus reported on in this paper cannot be publicly shared as it contains sensitive medical data.
compared to a 2.3% annual growth rate across all clinical notes, with or without ENDS documentation.

Our annotated corpus of ENDS mentions consisted of 4,351 annotations from the period 2010-2020. Table 1 shows a breakdown of the annotation types, with five top-level annotation categories and sixteen lower-level annotation categories. It can be seen that Active-User is the most frequent annotation type (43%) with Former-User the least frequent annotation type (3%). Of the Active-User group, the majority of mentions were related to tobacco cessation (either tobacco-cessation-related, cessation-counseling-pro or counseling-non-descriptive), with only a small proportion relating to attempts to quit ENDS. Four percent of mentions were found to be Irrelevant by annotators, with the majority of these Irrelevant posts referring to vaping cannabis. The initial annotation scheme development process yielded an agreement (kappa) of 0.7. The subsequent annotation effort yielded an overall agreement of 0.65 (0.85 for the top-level annotation categories Irrelevant, Usage-Unknown, Active-User, Non-User, and Former-User) on those portions of the corpus that were double annotated.

As discussed above, each ENDS mention is associated with patient demographic information derived from structured data in the VA EHR. Table 1 shows that — consistent with the wider VA patient population — 91% of annotated ENDS mentions are derived from the notes of male patients, and 9% of annotated ENDS mentions are derived from the notes of female patients. The difference between urban and rural locations was less marked than that between male and female, with 65% of annotated ENDS mentions derived from the notes of urban patients — or rather, patients who attended a VA facility in an urban area — and 35% derived from the notes of rural patients. Regarding age range, the majority of ENDS mentions were derived from individuals between the ages of 50 and 69, with the lowest number of mentions derived from the youngest age group (≤ 29). Documentation occurred in a range of clinical note types, with Other (a “catch-all” category that covers several internal medicine specialties) the most frequent note type (32%). Additionally, primary care notes and mental health notes were both well represented (31% and 12%, respectively).

### Table 1: Distribution of annotation categories (4,351 annotations in total)

<table>
<thead>
<tr>
<th>Annot. Type</th>
<th>Annot. Subtype</th>
<th>#Annotations</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Irrelevant</td>
<td>Total Annotations</td>
<td>155 (3.56%)</td>
<td>Vaping marijuana to help him with sleep.</td>
</tr>
<tr>
<td></td>
<td>Cannabis</td>
<td>86 (1.98%)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Other</td>
<td>69 (1.59%)</td>
<td>Used vaporizer when sick with flu</td>
</tr>
<tr>
<td>Usage-unknown</td>
<td>Total Annotations</td>
<td>1651 (37.95%)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cessation-counseling</td>
<td>382 (8.78%)</td>
<td>E-cigarettes may be bad for your health</td>
</tr>
<tr>
<td></td>
<td>Empty-template</td>
<td>13 (&lt; 1%)</td>
<td>Electronic cigarette: Y/N</td>
</tr>
<tr>
<td></td>
<td>Other</td>
<td>871 (20.02%)</td>
<td>Current tobacco user (excluding ecigs &amp; hookah</td>
</tr>
<tr>
<td></td>
<td>Non-exclusive</td>
<td>3858 (8.84%)</td>
<td>Current tobacco user (excluding ecigs &amp; hookah</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Tobacco/ECig: Y/N</td>
</tr>
<tr>
<td>Active-user</td>
<td>Total Annotations</td>
<td>1862 (42.79%)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Attempting-to-quit-ENDS</td>
<td>112 (2.57%)</td>
<td>Tapering off ecig</td>
</tr>
<tr>
<td></td>
<td>Tobacco-cessation-related</td>
<td>929 (21.35%)</td>
<td>Smoked tobacco but recently switched to vaping</td>
</tr>
<tr>
<td></td>
<td>Cessation-counseling-pro</td>
<td>12 (&lt; 1%)</td>
<td>Patient advised to continue ecig use</td>
</tr>
<tr>
<td></td>
<td>Counseling-non-descriptive</td>
<td>126 (2.90%)</td>
<td>Smoking ecigs: education provided</td>
</tr>
<tr>
<td></td>
<td>Non-descriptive</td>
<td>683 (15.70%)</td>
<td>Smokes ecigs</td>
</tr>
<tr>
<td>Non-user</td>
<td>Total Annotations</td>
<td>540 (12.41%)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Potential-user</td>
<td>135 (3.10%)</td>
<td>Pt planning on using ecig for cessation</td>
</tr>
<tr>
<td></td>
<td>Cessation-related</td>
<td>29 (&lt; 1%)</td>
<td>Pt not interested in using ecigs to quit</td>
</tr>
<tr>
<td></td>
<td>Passive-exposure</td>
<td>8 (&lt; 1%)</td>
<td>Girlfriend uses ecig</td>
</tr>
<tr>
<td></td>
<td>Non-descriptive</td>
<td>368 (8.46%)</td>
<td>Tobacco/ECig usage: No</td>
</tr>
<tr>
<td>Former-user</td>
<td>Total Annotations</td>
<td>143 (3.29%)</td>
<td>No longer vapes</td>
</tr>
</tbody>
</table>

**Discussion & Conclusion**

Our research has shown a considerable increase in ENDS documentation in the VA EHR over the period 2010-2020 (see Figure 2), with an average year-on-year growth rate of 53%, broadly reflecting ENDS use prevalence in the general population over the same period. In addition to the increase in the volume of ENDS mentions during the study period, our results show that the language of ENDS documentation has evolved over time, with terms like “vaping”
and “juul” (an ENDS brand name) supplanting previously popular terms like “electronic cigarette” and “e-cigarette” towards the end of the decade. This finding suggests that NLP systems designed to automatically identify ENDS use status must be flexible and adapt to change in language use over time. Unlike smoking status (i.e. combustible tobacco smoking status), where terminology is, if not standardized, then at least relatively stable over time (e.g. “2ppd”, “15 pack years”), ENDS documentation patterns are likely to continue to change quickly as technology and usage trends develop. As such, NLP systems designed to identify ENDS use must also evolve over time.

Winden et al.’s work on ENDS-related annotation scheme construction and corpus development provided a starting point for our VA-specific annotation scheme development process. However, there were some key differences in both the motivation of our work and in the specific characteristics of VA clinical note data that led us to substantially modify Winden et al.’s model for our specific needs. The work reported in this paper was primarily motivated by the goal of developing an annotation scheme and annotated corpus as a means of training and validating an NLP algorithm for identifying ENDS use status in the VA EHR. Given this motivation, it was necessary to include an Irrelevant annotation category for examples that include “false positive” ENDS-related keywords. Most of the examples of irrelevant mentions relate to cannabis use, particularly cannabis vaporization (e.g. “pt vapes marijuana”). A further distinctive feature of our annotation scheme is that it explicitly accounts for numerous types of semi-structured templates embedded in VA notes (e.g. “e-cigarette: Yes/No”) and use of facility-specific boilerplate language characteristic of VA notes (e.g. “the VA campus is a smoke free environment, including the use of e-cigarettes”). Excluding cannabis-related mentions, unfilled templates, and boilerplate language that does not directly refer to a patient’s specific ENDS use circumstances is a core requirement for an NLP algorithm capable of identifying ENDS use status at the VA.

<table>
<thead>
<tr>
<th>Annotation Subtype</th>
<th>Number of Annotations</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PATIENT GENDER</strong></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>3947 (91%)</td>
</tr>
<tr>
<td>Female</td>
<td>404 (9%)</td>
</tr>
<tr>
<td><strong>URBAN/RURAL</strong></td>
<td></td>
</tr>
<tr>
<td>Urban</td>
<td>2821 (65%)</td>
</tr>
<tr>
<td>Rural</td>
<td>1529 (35%)</td>
</tr>
<tr>
<td><strong>OUTPATIENT/INPATIENT</strong></td>
<td></td>
</tr>
<tr>
<td>Outpatient</td>
<td>3394 (78%)</td>
</tr>
<tr>
<td>Inpatient</td>
<td>949 (22%)</td>
</tr>
<tr>
<td><strong>AGE RANGE</strong></td>
<td></td>
</tr>
<tr>
<td>≤ 29</td>
<td>177 (4%)</td>
</tr>
<tr>
<td>30-49</td>
<td>977 (22%)</td>
</tr>
<tr>
<td>50-69</td>
<td>2,392 (55%)</td>
</tr>
<tr>
<td>≥ 70</td>
<td>805 (19%)</td>
</tr>
<tr>
<td><strong>NOTE TYPE</strong></td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>1383 (32%)</td>
</tr>
<tr>
<td>Primary care</td>
<td>1368 (31%)</td>
</tr>
<tr>
<td>Mental health</td>
<td>511 (12%)</td>
</tr>
<tr>
<td>Surgery</td>
<td>287 (7%)</td>
</tr>
<tr>
<td>Pharmacy</td>
<td>187 (4%)</td>
</tr>
<tr>
<td>Psychology</td>
<td>138 (3%)</td>
</tr>
<tr>
<td>Pulmonary disease</td>
<td>119 (3%)</td>
</tr>
<tr>
<td>Smoking cessation</td>
<td>77 (2%)</td>
</tr>
<tr>
<td>Emergency medicine</td>
<td>72 (2%)</td>
</tr>
<tr>
<td>Social work</td>
<td>69 (2%)</td>
</tr>
<tr>
<td>Cardiology</td>
<td>63 (1%)</td>
</tr>
<tr>
<td>Optometry</td>
<td>41 (1%)</td>
</tr>
<tr>
<td>Anesthesiology</td>
<td>36 (1%)</td>
</tr>
</tbody>
</table>

Table 2: Corpus characteristics
Research presented in this paper is not without limitations. First, our data collection spanned the period from January 1st 2010 to March 1st 2020 (i.e. immediately before the first major impacts of the COVID-19 pandemic in the United States were felt). Given systematic and drastic changes in ENDS documentation patterns at the VA since the onset of the pandemic and the increased scrutiny of potential ENDS-related COVID-19 transmission at the VA, it will likely be necessary to annotate additional contemporary ENDS mentions in order to ensure that a resulting NLP algorithm can be successfully applied to both contemporary and post-pandemic clinical notes. Second, our annotation scheme was designed primarily for the annotation of VA clinical notes, and may not be suitable without modification for other EHR environments. Third, given the relatively small size of the corpus compared to the universe of VA clinical notes and the stratified sampling method adopted in the corpus construction process, the corpus cannot in itself form a basis for forming epidemiological conclusions regarding, for example, ENDS use prevalence in VA patients.

In conclusion, the research described in this paper provides useful insights into changes in VA ENDS-related documentation practices during the period 2010-2020, as well as providing a resource for the future development and validation of NLP algorithms designed to identify ENDS use status.

Acknowledgments

We would like to take this opportunity to thank Mr Gregory Stoddard, MPH for the provision of valuable statistical advice that served to guide our sampling and stratification strategy.

Funding Statement

Research reported in this publication was partially supported by the National Institute on Drug Abuse of the National Institutes of Health under award number R03DA047577 and received further support in terms of resources and facilities from the Department of Veterans Affairs (VA) Informatics and Computing Infrastructure under award number VA HSR RES 13-457. The views expressed in this article are those of the authors and do not necessarily reflect the position or policy of the National Institutes of Health, the Department of Veterans Affairs, or the United States government.

Ethics Statement

The research reported on in this paper was approved by the University of Utah Institutional Review Board (IRB_00088382).

References


Testing of a Risk-Standardized Complication Rate Electronic Clinical Quality Measure (eCQM) for Total Hip and/or Total Knee Arthroplasty

Mica Curtin-Bowen, BA1, Troy Li, BS1, Avery Pullman, BS1, Alexandra Businger, MPH1, Stuart Lipsitz, ScD1,2, Ania Syrowatka, PhD1, Michael Sainlair, MS1, Tien Thai, BS1, Jay Lieberman, MD3, Aileen Davis, PhD4, Bonnie Blanchfield, ScD1,5, David W. Bates, MD, MSc1,2, Patricia C. Dykes, RN, PhD1,2;

1Brigham and Women’s Hospital, Boston, MA; 2Harvard Medical School, Boston, MA; 3Keck School of Medicine, The University of Southern California, Los Angeles, CA; 4University of Toronto, Ontario, CA; 5Harvard T.H. Chan School of Public Health, Boston, MA;

Abstract: Supported by the Centers for Medicare & Medicaid Services (CMS), Brigham and Women’s Hospital (BWH) has retooled the existing claims-based measures NQF1550 and NQF3493 into an electronic clinical quality measure (eCQM) to assess the risk-standardized complication rate (RSCR) following elective primary total hip (THA) and knee arthroplasty (TKA) at the clinician group level. This novel eCQM includes risk-adjustment for social determinants of health, includes all adult patients from all payers, leverages electronic health records (EHRs) rather than claims-based data, and includes both inpatient and outpatient procedures and complications which offers benefits compared to existing metrics. Following testing in two geographically different healthcare systems, the overall risk-standardized complication rate within 90 days following THA and TKA at the two sites was 3.60% (Site 1) and 3.70% (Site 2). This measure is designed for use in the Merit-Based Incentive Payment System (MIPS).

Introduction

Total hip arthroplasty (THA) and total knee arthroplasty (TKA) are common elective surgical procedures with 650,674 primary TKAs and 374,873 primary THAs performed in the United States in 2018 alone.1 Although rates of complications (including pneumonia, pulmonary embolism, periprosthetic joint infection, and wound infection) following these procedures are relatively low (ranging from 1.8%-9.0%)2 they can be life threatening.1 The danger of these complications and the variation in rates suggests a need for routine measurement and quality improvement.3 The demographics of patients receiving TKA and THAs are rapidly changing, and the fastest growing age group undergoing the procedures are individuals under the age of 65.4,5 Considering the projected rise in THA and TKA in the coming years,6 there is a need for robust, inclusive, and accessible evaluation measures for these procedures.7

Electronic clinical quality measures (eCQMs) are quality tools which are available to eligible hospitals, clinicians, or clinician groups that measure and track the quality of health care services.8 Starting in 2013, eCQMs across medical specialties have been developed to work within EHR systems to capture, export, calculate, and report information to provide outcome data without imposing documentation burden on clinicians.9 This proposed risk-standardized complication rate (RSCR) eCQM is designed for use within the Centers for Medicare & Medicaid (CMS) Merit-Based Incentive Payment System (MIPS) which provides an opportunity for clinicians and clinician groups to receive performance-based payment based on the quality of care provided to their patients and efforts taken to improve the quality of care.10

There are currently two related measures approved for use by the National Quality Forum (NQF), NQF155011 and NQF3493.12 Both measures are Clinical Quality Measures (CQMs) rather than eCQMs, meaning that they use claims-based data rather than electronic health record (EHR) data. These measures include only Medicare beneficiaries, and are limited to inpatient hospitalization complications.11,12,13 The objective of this research was to retool the NQF1550 and NQF3493 measures into an eCQM with an expanded inclusion population to provide a more meaningful way to analyze complication rates following joint replacement procedures.10 The
Testing of a Risk-Standardized Complication Rate eCQM for THA and TKA

retooled measure would include all adult patients and assess clinician group level complication rates in patients post-THA and TKA for use in the MIPS pathway.

This manuscript details the testing of the proposed RSCR eCQM in two geographically distant U.S. healthcare systems to assess post-THA/TKA complication rates at the clinician group level. If endorsed, this RCSR eCQM has the potential to improve patient outcomes as there are currently no NQF approved measures which track complications following elective primary THA and TKA that include all payers and all patients aged 18 years and older.

Materials and Methods

Study design

The development of this measure was comprised of alpha and beta testing of longitudinal data from all clinician groups within two large U.S. healthcare systems. This project was approved by the Human Subjects Committee, the Institutional Review Board (IRB) of Mass General Brigham. Alpha testing was used to assess the feasibility of implementing the measure into EHR systems, and beta testing quantitatively assessed the measure’s reliability and validity, providing RSCRs across sites, and additional information about feasibility.

The measure is comprised of a numerator and denominator, where the denominator is the total sample of THA patients and TKA patients who fit the inclusion criteria, and the numerator is the subset of patients who experienced the defined complications within 90 days of their procedure. This ratio is expressed as a percentage, where a lower RSCR is indicative of higher quality care. Complications are dichotomous as a yes/no event, where the occurrence of one or more complications results in inclusion of a patient in the numerator. Risk adjustment variables were harmonized with the NQF1550 to include: one or two procedures, comorbid conditions, and age. The model was then expanded to include race, sex, household income (using zip code as a proxy), language (English, non-English), smoking status, and body mass index (BMI). These risk adjustment variables were included in the retooled measure as they are all associated with worsened outcomes following THA and TKA; as this eCQM measures clinician level outcomes, the goal of the risk adjustment model was to reflect surgical performance without the influence of patient risk factors.

Post-procedural THA/TKA complications (harmonized with NQF1550) are defined as the following: acute myocardial infarction (AMI), pneumonia, sepsis/septicemia/shock, surgical site bleeding, pulmonary embolism, mechanical complications, periprosthetic joint infection/wound infection, and death. Table 1 below describes the number of days following the procedure date that a complication can be attributed to the procedure for inclusion in the numerator. A complication is counted in the numerator if it occurs at any point during the index admission, even if the index admission lasts longer than 7, 30, or 90 days. These time frames were harmonized with the NQF1550.

<table>
<thead>
<tr>
<th>Complication</th>
<th>Time Frame (days)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acute myocardial infarction, pneumonia, sepsis</td>
<td>7</td>
</tr>
<tr>
<td>Pulmonary embolism, surgical site bleeding, death</td>
<td>30</td>
</tr>
<tr>
<td>Wound infection/periprosthetic joint infection, mechanical comp.</td>
<td>90</td>
</tr>
</tbody>
</table>

Setting and Participants

eCQM testing used data from two large, geographically different, healthcare systems (described as ‘Site 1’ and ‘Site 2’) in the United States. Site 1 was comprised of six clinician groups with data from 2016-2019 and used the EHR vendor ‘Epic.’ Site 2 was comprised of eleven clinician groups with data from 2017-2019 and used the EHR vendor ‘Cerner.’ The following types of data were analyzed from EHR systems: administrative and billing data, patient demographics, procedures performed, and diagnoses.

All patients aged 18 years and older who received a primary elective THA and/or TKA from the two geographically different healthcare sites from 2016-2019 (Site 1) and 2017-2019 (Site 2) were included in the cohort. Harmonized with NQF1550, patients were excluded if they did not fit the inclusion criteria, if they left the hospital against medical advice (AMA), or if they had more than two THA/TKA procedure codes documented during the index admission. Unlike NQF1550 and NQF3493, both inpatient and outpatient procedures were analyzed. Clinician groups within each site who performed <50 THA and TKA procedures were excluded due to low sample size.

Alpha Testing

Alpha testing was performed to determine the feasibility of implementing the RSCR eCQM into EHR systems in two sites. Factors of interest in this testing included how eCQM data elements are captured, where they originate, how they are recorded and defined, their availability in different EHR systems, and if the eCQM data
element’s format and structure are consistent with the intent of the proposed measure. Alpha testing was performed using data from all patients aged 18 years and older that received a primary elective THA or TKA irrespective of inclusion/exclusion criteria.

The frequency of data elements in the samples from both sites was tested to confirm the data elements needed for risk-adjustment were routinely available. Reliability was tested using the NQF Feasibility Scorecard. Under CMS criteria, the feasibility of data elements is scored in the categories of data availability, data accuracy, data standards, and workflow. Chart reviews were conducted to assess measure agreement between the eCQM and manual review to confirm that the eCQM was correctly categorizing patients into the numerator, denominator, or excluded group. Separate chart reviews were performed on both random samples of patients who met the inclusion and exclusion criteria. Cases of disagreement between EHR abstraction and the eCQM were documented and discussed between team members. Disagreements were analyzed to implement changes that increased the accuracy of the eCQM. Kappa scores and 95% confidence intervals were assessed for each round of review.

**Beta Testing and Risk-Adjustment**

Beta testing was conducted using data from both sites to determine a complication rate by site, and by clinician groups within sites. Data from the included samples from both sites were randomly split into test and validation samples (50% in each sample) and fit the same generalized linear mixed model (with same covariates) in both samples. The demographics of the test and validation samples were compared (using P values) to explore similarities between the groups. Predicted over expected ratios (P/E) and 95% confidence intervals were calculated for the test and validation samples for all clinician groups who performed ≥50 procedures (for risk-adjustment, a minimum of 25 procedures were needed for both the test and the validation samples).

The RSCR is calculated as the ratio of the number of “predicted” to the number of “expected” cases with a complication within a given clinician group, multiplied by the unadjusted rate of the total sample. The measure estimates clinician group-level RSCRs using hierarchical logistic regression models. This approach models the log-odds of a patient experiencing a complication occurring within 90 days of the index admission using the risk adjustment variables and clinician group-specific intercepts. At the clinician group level, it models the group-specific intercepts as arising from a normal distribution. In this model, the clinician group intercept represents the underlying risk of complications for procedures performed by said group, after accounting for the patient risk factors in the risk adjustment model. If there were no differences among clinician groups, then after adjusting for patient risk, the group level intercepts should be identical across all groups.

The agreement between the estimates of the complication rates from each provider group in the test and validation samples was compared using a Spearman correlation coefficient. The risk-adjustment of this measure at the clinician group level was performed by multiplying the predicted/expected ratio in the test and validation samples by the total sample complication rate. 95% confidence intervals were calculated for each clinician group RSCR. Several statistical tests were performed to test the adequacy of the model. A C-statistic in the test and validation samples was calculated to see if discrimination was similar between both groups. A ROC curve was performed to compare the performance between test and validation samples. A Hosmer-Lemeshow test was used to assess the goodness of fit for the logistic regression model used for risk adjustment. An intraclass correlation coefficient (ICC) was performed to assess the variability in clinician groups.

**Results**

**Alpha Testing**

The total sample was comprised of 29,069 patients, 17,548 from Site 1, and 11,521 from Site 2. Across the total sample, data elements needed for risk adjustment were commonly populated (>98% complete) across both sites for insurance type, BMI, primary language, zip code, sex, admit age, and condition (Table 2). Data were missing for racial identity in Site 1 (95.67% complete) and smoking status in Site 2 (86.81% complete). Missing data comprised less than 3% for all other variables and therefore was viewed as unlikely to impact measure results.
Testing of a Risk-Standardized Complication Rate eCQM for THA and TKA

**Table 2: Frequency of Data Elements for Risk Adjustment in the Total Sample**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Site 1 (n=17600)</th>
<th>Site 2 (n=11521)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td># Missing</td>
<td># Available</td>
</tr>
<tr>
<td>Insurance Type</td>
<td>17</td>
<td>17583</td>
</tr>
<tr>
<td>BMI</td>
<td>119</td>
<td>17481</td>
</tr>
<tr>
<td>Primary Language</td>
<td>181</td>
<td>17419</td>
</tr>
<tr>
<td>Smoking Status</td>
<td>254</td>
<td>17346</td>
</tr>
<tr>
<td>Zip Code</td>
<td>22</td>
<td>17578</td>
</tr>
<tr>
<td>Sex</td>
<td>0</td>
<td>17600</td>
</tr>
<tr>
<td>Race</td>
<td>762</td>
<td>16838</td>
</tr>
<tr>
<td>Admit Age</td>
<td>0</td>
<td>17600</td>
</tr>
<tr>
<td>Condition</td>
<td>None of the cases had all NULL values for any of the condition columns</td>
<td></td>
</tr>
</tbody>
</table>

Using the NQF feasibility scorecard, the eCQM development team confirmed that all data elements were available, accurate, coded using nationally accepted terminology standards, and captured by clinicians during the course of care without additional documentation burden in both sites. Manual chart reviewers and the eCQM had excellent agreement, the included sample (n=217) had a final Kappa=0.95 (Table 3), and the excluded sample (n=217) had a final Kappa=1.0 (Table 4). These indicate near perfect inter-rater reliability in the included sample chart reviews, and perfect inter-rater reliability in the excluded sample chart reviews (Table 6).

**Table 3: Chart Reviews of the Included Sample (n=217)**

<table>
<thead>
<tr>
<th>CHART REVIEW ROUND</th>
<th>KAPPA</th>
<th>% AGREEMENT (95% Confidence Limits)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.833</td>
<td>88.89 (80.51 , 94.54)</td>
</tr>
<tr>
<td>2, 3, 4</td>
<td>0.864</td>
<td>91.07 (87.18 , 94.08)</td>
</tr>
<tr>
<td>5</td>
<td>0.950</td>
<td>96.67 (82.78 , 99.92)</td>
</tr>
</tbody>
</table>

**Table 4: Chart Reviews of the Excluded Sample (n=217)**

<table>
<thead>
<tr>
<th>CHART REVIEW ROUND</th>
<th>KAPPA</th>
<th>% AGREEMENT (95% Confidence Limits)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.855</td>
<td>93.33 (86.05 , 97.51)</td>
</tr>
<tr>
<td>2, 3, 4</td>
<td>0.879</td>
<td>95.19 (92.06 , 97.35)</td>
</tr>
<tr>
<td>5</td>
<td>1.000</td>
<td>100.00</td>
</tr>
</tbody>
</table>

**Beta Testing and Risk Adjustment**

At Site 1 from 2016-2019, 88.93% (n=15,607) of patients met the inclusion criteria with an unadjusted complication rate of 3.24%. At Site 2 from 2017-2019, 88.94% (n=10,247) of patients met the inclusion criteria with a crude complication rate of 4.28%. Descriptive statistics of the included sample can be found in Table 5.

**Table 5: Descriptive Statistics of the Included Sample from Site 1 (2016-2019) and Site 2 (2017-2019)**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Site 1</th>
<th>Site 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Admissions</td>
<td>17548</td>
<td>11521</td>
</tr>
<tr>
<td>Patients Included in the Measure</td>
<td>15607</td>
<td>10247</td>
</tr>
<tr>
<td>Patients Excluded from Measure</td>
<td>1941</td>
<td>1274</td>
</tr>
<tr>
<td>Number of Clinician Groups</td>
<td>6</td>
<td>11</td>
</tr>
<tr>
<td>Number of Inpatient/Outpatient Complications</td>
<td>506</td>
<td>439</td>
</tr>
<tr>
<td>Crude Complication Rate</td>
<td>3.24%</td>
<td>4.28%</td>
</tr>
</tbody>
</table>

**Demographics**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Site 1 (Mean (SD))</th>
<th>Site 2 (Mean (SD))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean Age (SD)</td>
<td>66.62 65.49</td>
<td></td>
</tr>
<tr>
<td>Age ≥ 65 years</td>
<td>59.60 56.52</td>
<td></td>
</tr>
<tr>
<td>18 ≤ Age ≤ 65 years</td>
<td>40.10 43.48</td>
<td></td>
</tr>
<tr>
<td>Male (%)</td>
<td>42.30 42.04</td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>90.00 68.54</td>
<td></td>
</tr>
<tr>
<td>Black/African American</td>
<td>3.68 10.62</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>2.06 20.77</td>
<td></td>
</tr>
<tr>
<td>English as first language</td>
<td>95.58 92.45</td>
<td></td>
</tr>
<tr>
<td>Smoker</td>
<td>5.28 5.83</td>
<td></td>
</tr>
<tr>
<td>Median Income</td>
<td>$85,935 $63,795</td>
<td></td>
</tr>
</tbody>
</table>

**THA/TKA Procedure**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Site 1</th>
<th>Site 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>THA procedure</td>
<td>44.31% 49.06%</td>
<td></td>
</tr>
<tr>
<td>TKA procedure</td>
<td>55.69% 50.94%</td>
<td></td>
</tr>
</tbody>
</table>

**Comorbid Conditions**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Site 1</th>
<th>Site 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decubitus ulcer or chronic skin ulcer</td>
<td>0.12% 0.08%</td>
<td></td>
</tr>
<tr>
<td>Dementia or other specified brain disorders</td>
<td>0.86% 1.51%</td>
<td></td>
</tr>
<tr>
<td>Diabetes mellitus (DM) or DM complications</td>
<td>11.96% 17.29%</td>
<td></td>
</tr>
<tr>
<td>Dialysis Status</td>
<td>0.12% 0.28%</td>
<td></td>
</tr>
<tr>
<td>Hemiplegia, paraplegia, paralysis, functional disability</td>
<td>0.66% 0.64%</td>
<td></td>
</tr>
<tr>
<td>Major psychiatric disorders</td>
<td>2.25% 1.23%</td>
<td></td>
</tr>
<tr>
<td>Metastatic cancer and acute leukemia</td>
<td>0.19% 0.06%</td>
<td></td>
</tr>
<tr>
<td>Morbid obesity</td>
<td>6.05% 11.35%</td>
<td></td>
</tr>
<tr>
<td>Osteoarthritis of hip or knee</td>
<td>99.26% 28.50%*</td>
<td></td>
</tr>
<tr>
<td>Osteoporosis and other bone/cartilage disorders</td>
<td>9.96% 5.81%</td>
<td></td>
</tr>
<tr>
<td>Other Congenital Deformity of Hip (joint)</td>
<td>0.58% 0.95%</td>
<td></td>
</tr>
<tr>
<td>Other major cancers</td>
<td>2.72% 1.96%</td>
<td></td>
</tr>
<tr>
<td>Other injuries</td>
<td>0.45% 0.04%</td>
<td></td>
</tr>
<tr>
<td>Pleural effusion/pneumothorax</td>
<td>0.08% 0.41%</td>
<td></td>
</tr>
<tr>
<td>Pneumonia</td>
<td>0.22% 0.00%</td>
<td></td>
</tr>
<tr>
<td>Post traumatic osteoarthritis</td>
<td>0.00% 0.37%</td>
<td></td>
</tr>
<tr>
<td>Protein-calorie malnutrition</td>
<td>0.07% 7.39%</td>
<td></td>
</tr>
<tr>
<td>Renal Failure</td>
<td>6.70% 0.20%</td>
<td></td>
</tr>
<tr>
<td>Respiratory/heart/digestive/urinary/other neoplasms</td>
<td>0.55% 4.72%</td>
<td></td>
</tr>
<tr>
<td>Rheumatoid arthritis &amp; inflammatory connective tissue disease</td>
<td>6.17% 0.70%</td>
<td></td>
</tr>
<tr>
<td>Stroke</td>
<td>0.06% 0.10%</td>
<td></td>
</tr>
</tbody>
</table>
Testing of a Risk-Standardized Complication Rate eCQM for THA and TKA

| Bone/joint/muscle infections/ necrosis | 3.42% | 3.19% | Major Complication of Medical care and trauma | 0.67% | 2.31% |
| Cardiorespiratory failure and shock | 0.73% | 1.27% | Vascular or circulatory disease | 9.35% | 9.30% |
| Coronary atherosclerosis or angina | 10.26% | 8.14% | Vertebral fractures without spinal cord injury | 0.01% | 0.11% |
| COPD | 4.49% | 3.55% | Trauma | 0.04% | 4.85% |

*Low rate of osteoarthristis in Site 2 is validated; Site 2 comorbid condition information was pulled from ICD10 codes upon discharge following THA and/or TKA, while Site 1 comorbid condition information was pulled from ICD10 codes upon admission prior to the THA and/or TKA.

Site 1 (n=15,607) had an overall unadjusted complication rate of 3.24% with rates ranging from 1.36%-4.51% across clinician groups. Site 2 (n=10,247) showed an overall unadjusted complication rate of 4.28%, with rates ranging from 3.02%-26.79% across groups. The total sample complication rate was 3.66%.

Table 6 displays the unadjusted complication rates with the associated test and validation samples of the predicted/expected ratios by clinician group, where 50% of patients per clinician group are in the test sample, and 50% of the patients per clinician group are in the validation sample. With these ratios, 1.0 represents the overall complication rate in the total sample, and the higher or lower the P/E score is, the higher or lower the complication rate within that specific group. Site 1 Group G and Site 2 Group A were excluded due to low sample size. A full list of the logistic regression coefficient estimates used to calculate the P/E can be found in Table 10.

<table>
<thead>
<tr>
<th>Site, Clinician Group</th>
<th>Unadjusted Rate</th>
<th>Test Sample</th>
<th>Validation Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sample Size</td>
<td>Test P/E</td>
<td>Test 95% CI</td>
</tr>
<tr>
<td>A</td>
<td>3.505</td>
<td>2853</td>
<td>1.036</td>
</tr>
<tr>
<td>B</td>
<td>1.524</td>
<td>328</td>
<td>0.933</td>
</tr>
<tr>
<td>C</td>
<td>3.265</td>
<td>1991</td>
<td>1.017</td>
</tr>
<tr>
<td>D</td>
<td>1.835</td>
<td>109</td>
<td>0.973</td>
</tr>
<tr>
<td>E</td>
<td>4.985</td>
<td>662</td>
<td>1.063</td>
</tr>
<tr>
<td>F</td>
<td>2.528</td>
<td>1859</td>
<td>0.884</td>
</tr>
<tr>
<td>G</td>
<td>2.667</td>
<td>75</td>
<td>0.992</td>
</tr>
<tr>
<td>H</td>
<td>11.688</td>
<td>154</td>
<td>1.153</td>
</tr>
<tr>
<td>J</td>
<td>25.926</td>
<td>27</td>
<td>1.042</td>
</tr>
<tr>
<td>K</td>
<td>3.777</td>
<td>1059</td>
<td>1.004</td>
</tr>
<tr>
<td></td>
<td>3.509</td>
<td>228</td>
<td>0.979</td>
</tr>
<tr>
<td></td>
<td>4.154</td>
<td>1661</td>
<td>1.029</td>
</tr>
<tr>
<td></td>
<td>6.849</td>
<td>73</td>
<td>0.983</td>
</tr>
<tr>
<td></td>
<td>4.000</td>
<td>425</td>
<td>0.958</td>
</tr>
<tr>
<td></td>
<td>4.854</td>
<td>206</td>
<td>1.031</td>
</tr>
<tr>
<td></td>
<td>4.608</td>
<td>217</td>
<td>0.969</td>
</tr>
<tr>
<td></td>
<td>3.021</td>
<td>993</td>
<td>0.958</td>
</tr>
</tbody>
</table>

We compared sociodemographic characteristics of patients included in our test and validation samples (Table 7) and found there were no differences at the patient level (p=0.434-0.840), by patient SES variables (p=0.085-0.820) or between clinician groups (p=0.999). The test and validation samples give similar rankings of the six clinician groups with respect to the predicted/expected ratios and gives a Spearman rank correlation of 0.978. The model has similar performance in both the test and validation samples (ROC=0.670 in both the test and validation sample).

<table>
<thead>
<tr>
<th>Variable [n (%)]</th>
<th>Test Sample</th>
<th>Validation Sample</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Site 1</td>
<td>7802 (60.39)</td>
<td>7805 (60.35)</td>
<td>0.945</td>
</tr>
<tr>
<td>Site 2</td>
<td>5118 (39.61)</td>
<td>5129 (39.66)</td>
<td></td>
</tr>
</tbody>
</table>

Table 7: Descriptive Statistics of Sample of Test and Validation Groups

<table>
<thead>
<tr>
<th>Number of Admissions by Group:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Site 1 A</td>
</tr>
<tr>
<td>Site 1 B</td>
</tr>
<tr>
<td>Site 1 C</td>
</tr>
<tr>
<td>Site 1 D</td>
</tr>
<tr>
<td>Site 1 E</td>
</tr>
<tr>
<td>Site 1 F</td>
</tr>
<tr>
<td>Site 2 B</td>
</tr>
<tr>
<td>Site 2 C</td>
</tr>
<tr>
<td>Site 2 D</td>
</tr>
<tr>
<td>Site 2 E</td>
</tr>
<tr>
<td>Site 2 F</td>
</tr>
<tr>
<td>Site 2 G</td>
</tr>
<tr>
<td>Site 2 H</td>
</tr>
<tr>
<td>Site 2 I</td>
</tr>
</tbody>
</table>

356
Testing of a Risk-Standardized Complication Rate eCQM for THA and TKA

<table>
<thead>
<tr>
<th>Healthcare Site, Clinician Group</th>
<th>Adjusted Complication Rate (%)</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Site 1 Overall Rate:</td>
<td>3.604%</td>
<td></td>
</tr>
<tr>
<td>1 A</td>
<td>3.784</td>
<td>3.13-4.44</td>
</tr>
<tr>
<td>1 B</td>
<td>3.433</td>
<td>2.59-4.28</td>
</tr>
<tr>
<td>1 C</td>
<td>3.693</td>
<td>3.0-4.39</td>
</tr>
<tr>
<td>1 D</td>
<td>3.547</td>
<td>2.67-4.43</td>
</tr>
<tr>
<td>1 E</td>
<td>3.902</td>
<td>3.13-4.68</td>
</tr>
<tr>
<td>1 F</td>
<td>3.267</td>
<td>2.59-3.94</td>
</tr>
<tr>
<td>1 G</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Site 2 Overall Rate:</td>
<td>3.694%</td>
<td></td>
</tr>
<tr>
<td>2 A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>2 B</td>
<td>3.662</td>
<td>2.76-4.57</td>
</tr>
<tr>
<td>2 C</td>
<td>4.249</td>
<td>3.38-5.11</td>
</tr>
<tr>
<td>2 D</td>
<td>3.832</td>
<td>3.15-4.51</td>
</tr>
<tr>
<td>2 E</td>
<td>3.687</td>
<td>2.95-4.42</td>
</tr>
<tr>
<td>2 F</td>
<td>3.579</td>
<td>2.79-4.37</td>
</tr>
<tr>
<td>2 G</td>
<td>3.750</td>
<td>3.07-4.42</td>
</tr>
<tr>
<td>2 H</td>
<td>3.579</td>
<td>2.84-4.32</td>
</tr>
<tr>
<td>2 I</td>
<td>3.521</td>
<td>2.78-4.26</td>
</tr>
<tr>
<td>2 J</td>
<td>3.783</td>
<td>2.93-4.64</td>
</tr>
<tr>
<td>2 K</td>
<td>3.530</td>
<td>2.79-4.27</td>
</tr>
<tr>
<td>2 L</td>
<td>3.464</td>
<td>2.74-4.18</td>
</tr>
</tbody>
</table>

Unadjusted complication rates demonstrated a large amount of variation between the two healthcare systems, as well as between individual clinician groups. In the unadjusted results, the crude rate at Site 2 (4.28%) was 1.32 times greater than the unadjusted complication rate at Site 1 (3.24%). When accounting for the risk factors, rates across sites became more similar. The overall Site 1 adjusted rate was 3.60% (ranging from 3.27%-3.90%). The overall Site 2 adjusted rate was 3.70% (ranging from 3.46%-4.25%) (Table 8).

Table 8: Risk-Adjusted Complication Rates with 95% Confidence Intervals

Unadjusted complication rates demonstrated a large amount of variation between the two healthcare systems, as well as between individual clinician groups. In the unadjusted results, the crude rate at Site 2 (4.28%) was 1.32 times greater than the unadjusted complication rate at Site 1 (3.24%). When accounting for the risk factors, rates across sites became more similar. The overall Site 1 adjusted rate was 3.60% (ranging from 3.27%-3.90%). The overall Site 2 adjusted rate was 3.70% (ranging from 3.46%-4.25%) (Table 8).
Testing of a Risk-Standardized Complication Rate eCQM for THA and TKA

Differences in patient characteristics were controlled well across both sites, with an intraclass correlation coefficient (ICC) of 0.0055 (95% CI = -0.017 - 0.027), meaning that only 0.55% of variation in complications was explained by provider-group after risk-adjustment. Patient level characteristics (within effects) were predictive of the outcome with a c-statistic of 0.672 (test: 0.674, validation: 0.670). A Hosmer-Lemeshow calibration approach was used to test the goodness of fit of the logistic regression model. The Hosmer-Lemeshow calibration approach resulted in a P value of 0.8202, where a P value >0.10 demonstrates good fit between the predicted and expected values.24

Fig. 1: Total Sample and Results

Discussion

The overall risk-adjusted complication rates following THA and TKA ranged from 2.84%-3.43% in Site 1, and from 3.0%-3.75% in Site 2. The increase in overall rate (3.604% in Site 1, 3.694% in Site 2) compared to the NQF1550 overall rate (2.4%)2 demonstrates the strength of the proposed measure in capturing a more accurate complication rate by expanding the inclusionary population of the measure to include both inpatient and outpatient procedures and complications.

Several limitations were acknowledged during the measure testing process. Data were used from two different sites, however additional testing at other health systems is needed to fully assess eCQM performance. Measure testing across sites is both time consuming and financially straining, and it is challenging to test an eCQM across multiple sites. Provider groups from both sites used in the analysis were affiliated with large academic medical centers, meaning that these rates do not reflect complication rates in smaller community health systems or independent clinician groups.

At the health system policy level, differences were seen in data element availability and diagnostic practices. Although smoking status information was only available for 86.81% in Site 2, less than 3% of data was missing across all other variables and was not likely to impact measure results. In the Site 1 sample, 99.26% of patients had a diagnosis of osteoarthritis, these high rates were to be expected as joint arthroplasties are recommended for patients who suffer from osteoarthritis.25 In comparison, only 28.50% of patients in the Site 2 sample had a diagnosis of osteoarthritis, reflecting differences in how these conditions are defined in EHRs. An explanation for these differences could be that Site 1 documented comorbid conditions upon index admission prior to the TJA procedure, and that Site 2 documented comorbid conditions following the TJA procedure. Rates of post-traumatic osteoarthritis were low in both sites (Site 1: 0.0%, Site 2: 0.37%). As this measure quantifies the complication rate among patients undergoing primary elective THA and TKAs, patients with a diagnosis of post traumatic osteoarthritis (which indicates a joint injury)26 are likely to be excluded as these procedures would no longer classify a procedure as ‘primary’ or ‘elective.’ This measure focuses on primary elective THA and TKA procedures because procedures associated with injury face higher risks of complications and mortality and are not equally reflective of clinician group performance.27

Several statistical tests were performed to assess the risk adjustment model, including a C-statistic of 0.672. A C-statistic of 0.5 indicates that the predictive model is no better than random chance, and a C-statistic of 0.7 or
Testing of a Risk-Standardized Complication Rate eCQM for THA and TKA

above indicates a good model, indicating that our model was better at predicting complications than random chance, but has room for improvement in future site testing. In comparison, the C-statistic for the NQF3493, which is already implemented into the MIPS pathway, was 0.65 across datasets.

The risk adjustment model for the RSCR eCQM was developed using published literature and stakeholder advisement from patients, healthcare plan payers, and providers. More information on the development of the risk-adjustment model can be found in Dykes et al., 2021. There are concerns within the field of quality measurement that payment programs, like MIPS, subject hospitals with higher proportions of socioeconomically disadvantaged populations to disproportionate financial penalties where providers are penalized for patient population risk factors, rather than clinical performance. During the measure development process, patient demographics including race, income level, and primary language were found to be associated with worsened outcomes following THA and/or TKA, and were included in the risk-adjustment model to quantify rates reflective of clinician group performance. While this risk-adjustment strategy aims to reflect surgical performance, measure developers must also acknowledge that adjusting for race and income may imply that worsened outcomes by race and income are justified or are not to be further explored. While the RSCR eCQM adjusts for socioeconomic variables, future research is needed to understand the significant differences in TJA outcomes by race and income, and to amend these disparities with equity-based interventions.

In comparison to the NQF1550 and NQF3493 measures, the RSCR eCQM provides several unique benefits. Firstly, this eCQM uses widely available and routinely collected EHR data, making it possible to use the eCQM in a timely manner without imposing documentation burden on clinicians. Using EHR data coded with established standards also means that this measure is interoperable between different EHR systems (e.g., Epic, Cerner) that use national coding standards. Secondly, including both inpatient and outpatient complications in the measure gives more insight into the true THA/TKA complication rate. Next, as this eCQM involves elective primary THA and TKA procedures from all payers, rather than solely from Medicare beneficiaries, the inclusion age is lowered from 65 to 18 years, which is especially relevant as the fastest growing segment of the population receiving these procedures are under 65 years of age. Finally, this proposed eCQM risk-adjusts for a more diverse set of health determinants than the existing NQF1550. As MIPS provides financial incentives to physicians to achieve lower complication rates, risk-adjusting for patient variables which are known to be associated with higher complication rates means physicians would not face MIPS penalties by treating complex patients. Table 9 describes the similarities and differences between the two existing measures and the proposed RSCR eCQM.

| Table 9: Comparison Between Existing RSCR Measures and the Proposed RSCR eCQM |
|---------------------------------|--|------------------|-----------------------|
| **Care Setting:**               | Hospital Level | Individual Clinician, Clinician Group Level | Clinician Group Level |
| **Measure Type**                | CQM (claims based) | eCQM (EHR based) |
| **Sample Population:**          | Medicare Fee for Service Beneficiaries | All Payers |
| **Insurance Type:**             | Patients age 65+ | Patients age 18+ |
| **Age:**                        | Inpatient Procedures | Inpatient and outpatient procedures and complications |
| **Location:**                   | Inpatient Procedures | Inpatient and outpatient procedures and complications |
| **Risk Adjustment Variables:**  | Age | Sex | Number of Procedures (1,2) | Procedure Type (THA, TKA) | Comorbid Conditions |
|                                | Age | Sex | Number of Procedures (1,2) | Procedure Type (THA, TKA) | Comorbid Conditions |
|                                | +Race | +Income (Zip Code) | +Smoking status | +Primary Language | +Body Mass Index (BMI) |

The RSCR eCQM has been submitted for NQF approval and CMS Measures Under Consideration (MUC) list consideration. If endorsed, this will be the only NQF approved eCQM which measures the risk-standardized complication rate following THA and TKA for all adult patients from all payers. If this measure is endorsed, the RSCR eCQM has the potential to enhance the MIPS program by informing healthcare provider groups about opportunities to improve care, strengthen incentives for quality improvement, and ultimately improve the quality of care received by patients.

References:

Testing of a Risk-Standardized Complication Rate eCQM for THA and TKA


Testing of a Risk-Standardized Complication Rate eCQM for THA and TKA


| Table 10: Risk-Adjustment Variables and the Estimated Logistic Regression Coefficients |
|------------------------------------------|-----------------|-----------------|
| Risk Adjustment Variables: Demographics, Location | Logistic Regression Coefficient Estimate | Risk Adjustment Variables: Comorbid Conditions | Logistic Regression Coefficient Estimate |
| Intercept | -4.3998 | Bone/joint/muscle infections/necrosis | 0.7509 |
| age | -0.02427 | Cardiorespiratory failure and shock | 1.1101 |
| age*age | 0.000231 | Chronic coronary atherosclerosis or angina | 0.2123 |
| MEDIAN INCOME 1000 | -0.00199 | COPD | 0.06342 |
| MEDIAN_IN*MEDIAN_INC | 0.00013 | Decubitus ulcer or chronic skin ulcer | 0.4694 |
| Private | -0.1036 | Stroke | 0.1201 |
| female | -0.0429 | Trauma | 0.9784 |
| hip | 0.2031 | Vascular or circulatory disease | 0.3407 |
| black | 0.5256 | Vertebral fractures without spinal cord injury | 0.4027 |
| white | 0.06731 | Dementia or other specified brain disorders | -0.04277 |
| English | -0.08302 | Diabetes mellitus (DM) or other DM complications | 0.09384 |
| smoke | 0.2132 | Dialysis status | -0.05276 |
| bmi | 0.05095 | Hemiplegia, paraplegia, paralysis, functional disability | 0.9639 |
| bmi *bmi | -0.00028 | Major psychiatric disorders | 0.6485 |
| Site 1 A | 0.03324 | Metastatic cancer and acute leukemia | -0.3755 |
| Site 1 B | -0.07227 | Morbid obesity | 0.2703 |
| Site 1 C | 0.01722 | Osteoarthritis of hip or knee | 0.1036 |
| Site 1 D | -0.02833 | Osteoporosis and other bone/cartilage disorders | 0.2003 |
| Site 1 E | 0.06363 | Other congenital deformity of hip (joint) | 0.3851 |
| Site 1 F | -0.1283 | Other major cancers | 0.6566 |
| Site 2 A | -0.00802 | Other injuries | -0.1296 |
| Site 2 B | 0.1599 | Pleural effusion/pneumothorax | 2.0178 |
| Site 2 C | 0.06122 | Pneumonia | 0.5992 |
| Site 2 D | 0.004694 | Post traumatic osteoarthritis | 0.4558 |
| Site 2 E | -0.02274 | Protein calorie malnutrition | 0.465 |
| Site 2 F | 0.0303 | Renal Failure | 0.1725 |
| Site 2 G | -0.02003 | Respiratory/heart/digestive/urinary/other neoplasms | 0.6375 |
| Site 2 H | -0.04814 | Rheumatoid arthritis & inflammatory connective tissue disease | 0.0702 |
| Site 2 I | 0.0329 | | |
| Site 2 J | -0.03547 | | |
| Site 2 K | -0.04651 | | |
Mapping the Read2/CTV3 controlled clinical terminologies to Phecodes in UK Biobank primary care electronic health records: implementation and evaluation

Spiros Denaxas\textsuperscript{1,3,4,5,6}, Ge Liu\textsuperscript{2}, Qiping Feng\textsuperscript{2}, Ghazaleh Fatemifar\textsuperscript{2,3}, Lisa Bastarache\textsuperscript{1}, Eric V. Kerchberger\textsuperscript{2}, Aroon D. Hingorani\textsuperscript{1,3,4}, Tom Lumbers\textsuperscript{1,3}, Josh F. Peterson\textsuperscript{2}, Wei-Qi Wei\textsuperscript{2}, Harry Hemingway\textsuperscript{1,3,4,6}

\textsuperscript{1}University College London, London, UK; \textsuperscript{2}Vanderbilt University Medical Center, Nashville, TN, USA; \textsuperscript{3}Health Data Research UK, London, UK; \textsuperscript{4}BHF Research Accelerator, London, UK; \textsuperscript{5}The Alan Turing Institute, London, UK, \textsuperscript{6}NIHR UCLH BRC, London, UK

Abstract

**Objective:** To establish and validate mappings between primary care clinical terminologies (Read Version 2, Clinical Terms Version 3) and Phecodes. **Methods:** We processed 123,662,421 primary care events from 230,096 UK Biobank (UKB) participants. We assessed the validity of the primary care-derived Phecodes by conducting PheWAS analyses for seven pre-selected SNPs in the UKB and compared with estimates from BioVU. **Results:** We mapped 92% of Read2 (n=10,834) and 91% of CTV3 (n=21,988) to 1,449 and 1,490 Phecodes. UKB PheWAS using Phecodes from primary care EHR and hospitalizations replicated all (n=22) previously-reported genotype-phenotype associations. When limiting Phecodes to primary care EHR, replication was 81% (n=18). **Conclusion:** We introduced a first version of mappings from Read2/CTV3 to Phecodes. The reference list of diseases provided by Phecodes can be extended, enabling researchers to leverage primary care EHR for high-throughput discovery research.

Introduction

Phenotype codes (Phecodes), are a hierarchical phenotype classification system offering a reference list of diseases for genetic and clinical research\textsuperscript{(1,2)}. Phecodes were originally created by mapping to coded electronic health records (EHR) using International Classification of Diseases (ICD) 9-Clinical Modification (CM), and more recently ICD-10 and ICD-10-CM\textsuperscript{(3)}. Phecodes may overcome some of the shortcomings of these and other clinical classification systems with each leaf node representing a clinically recognisable diagnosis. In the UK Biobank\textsuperscript{(4)}, a large scale resource of genomic sequencing and longitudinal phenotypic information, linkage has been obtained to longitudinal primary care information as well as information on hospital admissions. Although hospital episodes are coded using the ICD system, primary care data are coded using the Read version 2 (Read2) and Read version 3 (Clinical Terms 3 or CTV3) controlled clinical terminologies. Both terminologies provide a common standard vocabulary for clinicians and have been widely used for research\textsuperscript{(5,6)}.

Phecodes validated for the primary care setting, might offer several advantages. First, acute and chronic disease is often diagnosed within primary care settings in the UK. Patients with a disease treated in primary care may represent less severe forms of the disease than those who are treated in hospital. Phecodes can potentially act as data scaffolding providing a reference list of diseases between primary and secondary care settings. At a practical level, there is a need to reduce dimensionality of primary care EHR (with ~ 400K unique ontology concepts utilized) in a reproducible fashion to enable high-throughput phenotypic and genetic analyses\textsuperscript{(2)}. Controlled clinical terminologies that are used in primary care data however have not been mapped to Phecodes. As a result, the extent to which using primary care derived Phecodes offer a valid approach or whether they can provide additional statistical power or markers of disease severity is yet to be explored.

In this paper, we sought to develop and validate a mapping of clinical terminologies used in UK primary care EHR to Phecodes. Specifically, this was achieved through three objectives:

1. Create mappings between the Read controlled clinical terminology (Read2 and CTV3) and Phecodes.
2. Translate primary care EHR data in the UK Biobank to Phecodes.
3. Evaluate the mapping quality by conducting PheWAS analyses in the UK Biobank and BioVU and seeking to replicate previously-reported genotype–phenotype associations.
**Methods**

**Population and data sources**
We used data from the UK Biobank (UKB), a prospective study of 500,000 deeply phenotyped individuals recruited from England, Scotland and Wales. Longitudinal follow up for UKB participants is achieved through linkages to national data sources covering primary care, hospitalizations, cancer registrations and mortality. Specifically, information on hospital admissions for the entire cohort is available for all admitted patient episodes and is coded using ICD-9 and ICD-10 for (primary and secondary) diagnoses and OPCS-4 for surgical procedures. Approx 50% of participants have their primary care record linked which provides information on diagnoses, symptoms, laboratory results, referrals, examination findings and prescriptions.

Primary care EHR were extracted from four data sources based on three EHR vendors (two in England from Vision and TPP SystmOne, and two in Scotland and Wales combining data from EMIS and Vision). The basic unit of interaction in primary care is a consultation (similar to an admission in secondary care) during which the clinician records information directly into the EHR. In the UK, primary care healthcare concepts (except prescriptions) are recorded using Read codes, a hierarchical controlled clinical terminology. Read codes are used in two versions: Read Version 2 (Read2), also known as the 5-byte Read consists of ~100,000 concepts and Clinical Terms Version 3 (CTV3), consists of ~390,000 concepts and contains all Read 2 concepts. Both versions of Read are organized in 30 top level chapters in a fashion similar to ICD-10; e.g. chapter “G” is for Circulatory System Diseases Relationships in Read2 are derived directly through the code structure meaning only single parent-child term relationships are supported. In CTV3, relationships are defined through a separate relationship table and supports polyhierarchical relationships (e.g. ‘Infective pneumonia’ has two parents: ‘Pneumonia’ and ‘Acute lower respiratory tract infection’). SNOMED-CT, which is now becoming an international standard for recording information across healthcare settings and has >400K concepts encapsulates by definition all CTV3 terms since CTV3 was one of its core components.

**Mapping Read2 and CTV3 to Phecodes**
We implemented a systematic data-driven approach leveraging terminology reference sets, cross-map files, and expert input to map primary care EHR to Phecodes (Figure 1). The approach, which was similar for both Read2 and CTV3 is described below:

1. Retrieve all unique Read2/CTV3 codes from the UKB `gp_clinical` file.
2. Establish eligible codes Read2/CTV3 terms for mapping
   a. Remove Read2/CTV3 codes related to drug prescription events (defined in the Read Codes Drug and Appliance Dictionary (DAAD) dictionary and identified by having a lower-case first character).
   b. Remove invalid Read2/CTV3 codes (e.g. ‘XXXX’, ‘@@A2’) or local General Practitioner (GP) codes which could not be identified in the NHS terminology reference and are often used as placeholders (e.g. ‘UTEST’ or ‘UAB00’).
   c. Remove Read2/CTV3 codes related to occupations, examinations, symptoms, diagnostic and laboratory procedures (all codes in Read chapters 0 to 9 inclusive).
3. Map Read2/CTV3 codes to ICD-10 using the NHS Digital Technology Reference data Update Distribution (TRUD) (7) cross-map file and remove codes which were not not found in the map file.
   a. Remove Read2/CTV3 codes where the mapped ICD-10 chapter is X (Other external causes of accidental injury, self-harm), Y (Assault, legal interventional, complications) or Z (Factors influencing health status and contact with health services) as they were excluded in the original Phecode definitions.
   b. Exclude Read2/CTV3 codes mapping to an entire ICD-10 chapter (e.g. “H…. Respiratory diseases” maps to ICD-10 Chapter X “Diseases of Respiratory System”) since they are rarely used and are too broad.
4. Apply manual refinements e.g. recode specific ICD-10 leaf nodes to parent nodes (described in next paragraph)
5. Map ICD-10 codes to Phecodes using the existing translation file (3) (version 1.2beta).

Mappings between Read codes and ICD-10 defined in the “cross map” file used in the algorithm are defined in a different approach as outlined in Table 1. Where multiple target ICD-10 concepts existed (e.g. “one-to-block” mappings in Read2 or “type A” mappings in CTV3) we chose the broadest ICD-10 concept available (often, but not always, identified by the fourth digit being equal to ‘9’ e.g. L21.9 – Seborrheic dermatitis, unspecified). This was
done in order to avoid assigning a specific diagnosis when the source Read concept was broader than the candidate target ICD-10 concept. Where asterisk, dagger or asterisk and dagger terms were specified in the map file, we retained the most specific of ICD-10 (e.g. “H36.0 Diabetic Retinopathy” was preferred over “E14 Unspecified diabetes mellitus”). In instances where one Read2/CTV3 to many target ICD-10 concepts were defined in the map file, we added all individual ICD-10 codes as new records (e.g. Read2 code “F4K0.0 Scleritis and episcleritis” was mapped to two ICD-10 codes “H15.0 Scleritis” and “H15.1 Episcleritis”). In all other cases where a one-to-one map type was supplied, we used the target ICD-10 code specified. Finally, in order to align the target ICD-10 codes with the ICD-10-Phecode map, in some cases we upcoded (specific to broad) or downcoded (broad to specific) terms. Specifically, in some cases four character ICD-10 codes were remapped to children terms (e.g. “J34X Other disorders of nose and nasal sinuses” was remapped to “J34.8 Other specified disorders of nose and nasal sinuses”). In other cases, specific children terms were remapped to broader parent terms (e.g. “I48.9 Atrial fibrillation and flutter, unspecified” was remapped to “I48 Atrial fibrillation and flutter”).

Table 1. Map types specified by the NHS TRUD terminology cross-map files between source terms in Read2 and CTV3 to target ICD-10 codes. The Type column denotes the cardinality of the map; for example, one-to-one denotes one source term with exactly one target term while one-to-many denotes one source term with multiple potential target terms.

<table>
<thead>
<tr>
<th>Type</th>
<th>Example: Read V2/CTV3 =&gt; ICD-10 code</th>
<th>N (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Read2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>one to one</td>
<td>One source code can be mapped to exactly one ICD-10 code.</td>
<td>11,079</td>
</tr>
<tr>
<td></td>
<td>G5730 Atrial Fibrillation =&gt; I48 Atrial fibrillation and atrial flutter</td>
<td></td>
</tr>
<tr>
<td>one to block</td>
<td>One source code can be mapped to any one code from a specific ICD-10 block.</td>
<td>334</td>
</tr>
<tr>
<td></td>
<td>G30..Acute myocardial infarction =&gt; I21.0-I21.9</td>
<td></td>
</tr>
<tr>
<td>one to one of many</td>
<td>One source code can be mapped to one of multiple potential ICD-10 codes.</td>
<td>230</td>
</tr>
<tr>
<td></td>
<td>AB20. Candidiasis of mouth and oesophagus =&gt; B37.0 Candidal stomatitis or B37.8</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Candidiasis of other sites</td>
<td></td>
</tr>
<tr>
<td>asterisk</td>
<td>Asterisk codes relate to disease manifestations (as opposed to disease aetiology)</td>
<td>57</td>
</tr>
<tr>
<td></td>
<td>Cyu4K [X]Disorders of adrenal glands =&gt; E35.1* Disorders of adrenal glands</td>
<td></td>
</tr>
<tr>
<td>asterisk &amp; dagger</td>
<td>This map type combines both asterisk and dagger codes.</td>
<td>49</td>
</tr>
<tr>
<td></td>
<td>A053. Amoebic liver abscess =&gt; A06.4† Amoebic liver abscess (K77.0*) K77.0* Liver disorders in infectious and parasitic diseases classified elsewhere</td>
<td></td>
</tr>
<tr>
<td>dagger</td>
<td>Dagger codes relate to disease aetiology.</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td>N042. Other rheumatoid arthropathy =&gt; M05.3† Rheumatoid arthritis with involvement of other organs and systems</td>
<td></td>
</tr>
<tr>
<td>one to many</td>
<td>One source code maps to multiple ICD-10 codes that need to be combined</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>SP08Z Thrombosis of artery of transplanted kidney =&gt; N280 Ischemia and infarction of kidney + Z940 Kidney transplant status</td>
<td></td>
</tr>
<tr>
<td>CTV3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>Generic mapping, target ICD-10 code broader than source concept.</td>
<td>11,425</td>
</tr>
<tr>
<td></td>
<td>X101u Late onset asthma =&gt; J45.9 Asthma, unspecified</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>Default mapping - most acceptable amongst alternatives given absence of other information</td>
<td>11,064</td>
</tr>
<tr>
<td></td>
<td>G5730 Atrial fibrillation =&gt; I48.9 Atrial fibrillation and atrial flutter</td>
<td></td>
</tr>
</tbody>
</table>

364
Assigning Phecodes, separately in primary care and hospitalisation data

All unique codes from primary care (Read2, CTV3) and hospitalisations (ICD-10) from each UKB participant were extracted and translated into Phecodes. For hospitalization data, we included all ICD-10 codes (recorded as either a primary or secondary diagnosis) and used an existing ICD-10-Phecode translation file (3). In order to enhance the specificity and positive predictive value, individuals were considered a “case” if they had at least two occurrences (irrespective of when) of an ICD-10 code that mapped to a Phecode. Individuals were marked as “controls” if they did not have any ICD-10 codes belonging to the exclusion codes defined in the Phecode definition files. Individuals with ICD-10 codes in the Phecode exclusion range were excluded from the respective analyses. We excluded Phecodes that occurred in less than 100 participants (0.05 prevalence) in order to reduce data sparsity.

Comparative PheWAS analysis

To evaluate the quality of the Phecode data derived from primary care EHR, we performed a PheWAS analysis in the UK Biobank seeking to replicate existing known GWAS associations identified through previous analyses (8,9) in seven Single Nucleotide Polymorphisms (SNPs): rs3135388, rs17234657, rs2200733, rs1333049, rs6457620, rs8050136, and rs7903146. We obtained up-to-date estimates between the seven SNPs and 25 phenotypes which we used as comparators by conducting a PheWAS analysis in BioVU(10), the Vanderbilt University Medical Center biobank.

The main analysis used phenotypes derived from both primary care and hospitalization billing data in the UK Biobank. We performed a subsequent sensitivity analysis by analyzing phenotype data derived only from primary care EHR. SNP-phenotype associations were assessed using a logistic regression model assuming an additive genetic model. UK Biobank models were adjusted for sex, genotyping array and the first 11 principal components. BioVU models were adjusted for age at most recent medical encounter, sex, 10 principal components and length of EHR (defined as the time between each patient’s first and most recent medical encounter). We calculated and reported the expected effect sizes as odds ratios (OR) and assumed a SNP-phenotype association was significant if it’s P-value surpassed Bonferroni significance.

UK Biobank and Bio VU Genotyping and Quality Control (QC)

487,409 UKB participants were genotyped using one of two custom genome-wide arrays and data were imputed to a combination of the UK10K, 1000 Genomes Phase 3 and the Haplotype Reference Consortium (HRC) reference panels resulting in 93,095,623 variants. We applied additional variant level QC and excluded variants with: a) Fisher’s exact test <0.3, b) minor allele frequency (MAF) <1% and, c) a missing call rate of ≥5%. We applied individual-level QC and excluded participants with: a) excessive or minimal heterozygosity, b) more than 10 putative third-degree relatives as per the kinship matrix, c) no consent to extract DNA, d) sex mismatches between self-reported and genetic sex, and d) non-European ancestry (more details on QC provided elsewhere (11)). BioVU samples were genotyped on the Infinium Multi-Ethnic Genotyping Array (MEGA). To curate the genotyping data, a quality control step was conducted. Briefly, the quality control has been performed with following steps: a) the samples with per-indvividual call rate less than 95% were removed, b) the samples with wrongly assigned sex were deleted, c) the samples from related individuals (PI_HAT ≥ 0.25) were removed, d) other unexpected duplications were eliminated. A genome imputation process was applied to increase coverage of the GWAS, using the Michigan Imputation Server(12) and referring to the HRC.

Implementation

Phenotypic data extraction and translation was performed in Python 3.7.7, genetic data extraction and QC was performed using bgenix, PLINK 1.9 and qctool v2. PheWAS analyses were conducted in R v. 4.0.0 using the PheWAS library (https://github.com/PheWAS/PheWAS(. The Read-Phecode map file, programming scripts, and

<table>
<thead>
<tr>
<th>concepts exist.</th>
<th>G5730 Atrial fibrillation =&gt; I48.0 Paroxysmal atrial fibrillation, I48.1 Persistent atrial fibrillation, I48.2 Chronic atrial fibrillation</th>
</tr>
</thead>
<tbody>
<tr>
<td>E</td>
<td>Exact one to one mapping, no alternatives exist.</td>
</tr>
<tr>
<td>H322</td>
<td>Centrilobular emphysema =&gt; J43.2 Centrilobular emphysema</td>
</tr>
<tr>
<td></td>
<td>1,392 6%</td>
</tr>
</tbody>
</table>
documentation files will be made available under an open source license at the PheWAS catalogue(13) and the HDR UK Phenotype Portal(14).

Results
In the UKB, we processed 123,662,421 primary care clinical events: 3,6175,235 using Read2 and 87,487,186 using CTV3. We extracted 38,228 unique Read2 terms and 80,994 unique CTV3 terms which were recorded among 230,096 participants. After applying the mapping algorithm (described previously), we identified 11,775 and 23,881 unique Read2 and CTV3 terms respectively which were eligible for mapping. We mapped 10,834 (92%) of Read2 terms to 1,449 Phecodes, and 21,988 (91%) of CTV3 terms to 1,490 unique Phecodes (Table 2). We additionally processed 3,541,618 admitted patient hospitalization events from 435,632 patients and translated 6,758 unique ICD-10 primary and secondary diagnoses from 9,493,039 hospitalizations to Phecodes.

Table 2. Translation of Read Version 2 and Clinical Terms Version 3 terms to Phecodes.

<table>
<thead>
<tr>
<th></th>
<th>Read Version 2</th>
<th>Clinical Terms Version 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Events</td>
<td>36,175,235</td>
<td>87,487,186</td>
</tr>
<tr>
<td>Unique codes</td>
<td>38,228</td>
<td>80,994</td>
</tr>
<tr>
<td>Codes eligible for mapping</td>
<td>11,775</td>
<td>23,881</td>
</tr>
<tr>
<td>Unique Read codes mapped (%)</td>
<td>10,834 (92%)</td>
<td>21,988 (91%)</td>
</tr>
<tr>
<td>Unique ICD-10 codes</td>
<td>4,655</td>
<td>5,407</td>
</tr>
<tr>
<td>Unique Phecodes</td>
<td>1,449</td>
<td>1,490</td>
</tr>
</tbody>
</table>

In Read2, the top five unmapped terms were: a) G84. Hemorrhoids (n=9,601), b) Eu32z Depressive episode (n=7,702), unspecified, c) Eu32. Depressive episode (n=3,684), d) G843. External haemorrhoids, simple (n=1701), and e) N224. Ganglion and cyst of synovium, tendons, bursa (n=1,553). In CTV3, the most common unmapped terms were: a) XE2q5 Serum creatinine level (n=1,194,627), b) XE2q0 Serum sodium level (n=1,142,496), c) XE2pz Serum potassium level (n=1,140,859), d) XM0lt Serum urea level (n=1,041,244) and e) XE2eA Serum albumin level (n=1,014,201). In Read2, laboratory test results are in chapter 4 while in CTV3 concepts on laboratory measurements exist across multiple chapter chapters (e.g. “44M4. Serum albumin” and “XE2eA Serum albumin”).

The five most commonly recorded leaf Phecodes in hospital administrative data were: a) 401.10 Essential hypertension (n=100,940), b) 272.11 Hypercholesterolemia (n=35,033), c) 716.90 Arthropathy NOS (n=32,345), d) 550.20 Diaphragmatic hernia (n=21,203) and e) 562.10 Diverticularis (n=20,763). In primary care EHR data, the most frequently-recorded leaf Phecodes were: a) 745.0 Pain in joint (n=46,343), b) 401.1 Essential hypertension (n=40,645), c) 716.9 Arthropathy Not Elsewhere Specified (NOS) (n=30,516), d) 519.8 Other diseases of respiratory system, Not Elsewhere Classified (NEC) (n=29,146) and e) 760.0 Back pain (n=25,562).

In the primary analyses (Figure 2., Table 3) using Phecodes from hospitalization and primary care events, we performed a PheWAS in 408,415 UKB participants (54.4% female, mean age 59 SD 7.96) and 65,561 BioVU participants (55.6% female, mean age 57.87 SD 22.78) using 1,851 Phecodes respectively. We replicated 22 (100%) of 22 previously-reported genotype-phenotype associations with adequate statistical power across both of the datasets. In the secondary analyses, we performed a PheWAS in 185,648 UKB participants (~50% of the cohort is linked to primary care EHR, 54.5% female, mean age 58 SD 7.96) and 1,851 Phecodes and replicated 18 (81%) of 22 previously-reported PheWAS associations.
Figure 1. Exemplar flowchart of the process of mapping Read 2 and CTV3 terms, to ICD-10 and to Phecodes. CTV3 contains all Read2 terms but these were not duplicated in the display. CTV3=Clinical Terms Version 3.

Figure 2. Manhattan plot of phenome-wide association analyses in 185,648 participants (UKB, left) and 65,561 participants (BioVU, right) and 1,851 Phecodes derived from electronic health records. The red line shows the Bonferroni level of significance. Only phenotypes that cross the Bonferroni level of significance are annotated.
Table 3: PheWAS comparison of Phecodes derived from UK Biobank electronic health records (primary care EHR exclusively and from primary care and hospitalization data combined) to BioVU PheWAS analyses for seven variants with known associations. For each variant, the mapped genes and Genome Aggregation Database (gnomAD) identifiers are provided. P-Value columns with gray shading indicate results did not survive Bonferroni correction. Phenotypes with no identifier (in first column) are reported to illustrate differences in coding between data sources.

OR = Odds Ratio, PheWAS = Phenome-Wide Association Study. “n/a” denotes analyses which were not run due to a very low number of cases or no of cases due to a Phecode not being present in the data.

<table>
<thead>
<tr>
<th></th>
<th>Phenotype (phecode)</th>
<th>UK Biobank (all sources)</th>
<th>UK Biobank (primary care)</th>
<th>BioVU</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cases</td>
<td>OR</td>
<td>P Value</td>
<td>Cases</td>
</tr>
<tr>
<td>#</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Multiple sclerosis (335)</td>
<td>1486</td>
<td>2.42</td>
<td>2.79E-105</td>
</tr>
<tr>
<td>2</td>
<td>Type 1 diabetes (250.1)</td>
<td>2538</td>
<td>0.45</td>
<td>1.90E-48</td>
</tr>
<tr>
<td>3</td>
<td>Hypothyroidism NOS (244.4)</td>
<td>19943</td>
<td>0.82</td>
<td>3.33E-35</td>
</tr>
<tr>
<td>4</td>
<td>Celiac disease (557.1)</td>
<td>2056</td>
<td>0.64</td>
<td>6.06E-18</td>
</tr>
<tr>
<td></td>
<td>rs17234657 (AC108105.1, gnomAD: 5_40401407_T_G)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Regional enteritis (555.1)</td>
<td>1639</td>
<td>1.34</td>
<td>3.01E-10</td>
</tr>
<tr>
<td></td>
<td>rs2200733 (PITX2, gnomAD: 4_110789013_C_T)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Cardiac dysrhythmias (427.0)</td>
<td>30440</td>
<td>1.35</td>
<td>2.68E-124</td>
</tr>
<tr>
<td>7</td>
<td>Atrial fibrillation (427.21)</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td></td>
<td>Atrial flutter (427.22)</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td></td>
<td>Atrial fibrillation and flutter (427.2)</td>
<td>20962</td>
<td>1.53</td>
<td>2.78E-185</td>
</tr>
<tr>
<td></td>
<td>rs1333049 (CDKN2B-AS1, gnomAD: 9_22125504_G_C)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Coronary atherosclerosis (411.4)</td>
<td>16934</td>
<td>1.25</td>
<td>3.02E-88</td>
</tr>
<tr>
<td>9</td>
<td>Hyperlipidemia (272.1)</td>
<td>33679</td>
<td>1.06</td>
<td>1.14E-12</td>
</tr>
<tr>
<td>10</td>
<td>Angina pectoris (411.3)</td>
<td>15164</td>
<td>1.19</td>
<td>1.78E-48</td>
</tr>
<tr>
<td></td>
<td>Condition</td>
<td>Sample Size</td>
<td>OR</td>
<td>p-value</td>
</tr>
<tr>
<td>---</td>
<td>----------------------------------</td>
<td>-------------</td>
<td>------</td>
<td>---------</td>
</tr>
<tr>
<td>11</td>
<td>Unstable angina (411.1)</td>
<td>3458</td>
<td>1.19</td>
<td>4.64E-13</td>
</tr>
<tr>
<td>12</td>
<td>Myocardial infarction (411.2)</td>
<td>13755</td>
<td>1.2</td>
<td>8.80E-51</td>
</tr>
<tr>
<td>13</td>
<td>Other chronic IHD, unspecified (411.8)</td>
<td>16934</td>
<td>1.18</td>
<td>6.78E-51</td>
</tr>
<tr>
<td></td>
<td>rs6457620 (HLA-DQB1, gnomAD: 6_32696222_G_C)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Rheumatoid arthritis (714.1)</td>
<td>5006</td>
<td>1.57</td>
<td>3.96E-106</td>
</tr>
<tr>
<td>15</td>
<td>Multiple sclerosis (335.0)</td>
<td>1486</td>
<td>0.63</td>
<td>3.14E-33</td>
</tr>
<tr>
<td>16</td>
<td>Celiac disease (557.1)</td>
<td>2056</td>
<td>0.27</td>
<td>3.07E-248</td>
</tr>
<tr>
<td></td>
<td>rs8050136 (FTO, gnomAD: 16_53782363_C_A)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>Obesity (278.1)</td>
<td>11088</td>
<td>1.15</td>
<td>6.60E-24</td>
</tr>
<tr>
<td>18</td>
<td>Type 2 diabetes (250.2)</td>
<td>26517</td>
<td>1.11</td>
<td>8.15E-28</td>
</tr>
<tr>
<td>19</td>
<td>Essential hypertension (401.1)</td>
<td>100106</td>
<td>1.04</td>
<td>4.63E-12</td>
</tr>
<tr>
<td></td>
<td>Obstructive sleep apnea (327.32)</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>20</td>
<td>Sleep apnea (327.3)</td>
<td>4890</td>
<td>1.12</td>
<td>1.98E-08</td>
</tr>
<tr>
<td></td>
<td>rs7903146 (TCF7L2, gnomAD: 10_112998590_C_T)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>Type 2 diabetes (250.2)</td>
<td>26517</td>
<td>1.34</td>
<td>3.48E-210</td>
</tr>
<tr>
<td>22</td>
<td>Type 2 diabetes ophthalmic compl. (250.23)</td>
<td>1494</td>
<td>1.34</td>
<td>2.46E-14</td>
</tr>
</tbody>
</table>
Discussion

In this study, we described the process of mapping primary care EHR data recorded using the Read2 and CTV3 controlled clinical terminologies to Phecodes. We performed a PheWAS in two contemporary international biobanks, the UK Biobank (UK) and BioVU (US) and provided evidence towards the validity of Phecodes derived from longitudinal primary care EHR by showing concordant findings. Similar to the US, UK hospital data are collected for billing purposes which may influence the coding process however billing codes in BioVU additionally include ambulatory care. UK primary care has a true EHR generated and captured during clinical care by healthcare professionals; billing and funding may also have a modest effect (e.g. through the Quality and Outcomes Framework (15)). Reassuringly, the distribution of most frequently mapped Phecodes between sources reflected these differences with Phecodes on symptoms (e.g. back pain, joint pain) being more prevalent in primary care EHR.

Our study is the first, to our knowledge, to develop and evaluate Phecodes sourced from primary care EHR data and show the validity of the translation file through a PheWAS analyses in two international resources. Comparison between UKB and BioVU uncovered challenges with regards to the resolution of diagnosis codes used in the UK and the US. ICD-10-CM used in the US has a higher fidelity than ICD-10 used in the UK which directly influenced the level at which Phecodes were assigned. For example, in BioVU, the availability of ICD-10-CM codes meant that sleep apnea was recorded using a lower-level Phecode “Obstructive sleep apnea (327.32)” whereas in the UK, a higher level one “Sleep apnea (327.3)”. As a result, while the PheWAS analyses failed to show concordant results between the resources, each individual analyses on the Phecodes that were actually available in the data sources was significant. The same effect was observed in atrial fibrillation and flutter which are recorded using a single ICD-10 code in UK data but are split across two different codes in ICD-10-CM (and as a result two different Phecodes).

The main limitation of our study is the fact that only approximately 50% of the population in the UK Biobank has primary care data available. As a result, our PheWAS analyses using Phecodes derived from primary care EHR failed to replicate previous associations due to low statistical power (e.g. 740 cases of sleep apnoea recorded in primary care EHR). Some diagnoses in primary care were not mapped to Phecodes. This could be due to the fact that the target ICD-10 code was not in the ICD-10-Phecode map; for example, in the case of haemorrhoids we observed a mismatch between the ICD-10 code available in the cross-map file “K64 Hemorrhoids and perianal venous thrombosis” and the ICD-10 code available in the ICD-10-Phecode map “I84 Haemorrhoids”. Another reason for missed mappings was that source Read2/CTV3 concepts were too generic (e.g. S3zFracture of unspecified bones) and could not be mapped. In some cases, such as Read code Eu32z "[X]Depressive episode, unspecified", the target term was ICD-10 F32 "Major depressive disorder, single episode" which is a non-billable code with no associated Phecode. Finally, some Read codes (e.g. E2273 Impotence) were not defined in the Read-ICD-10 cross-map files and as a result were not mapped to a relevant Phecode during data extraction.

Despite the relatively small number of events not mapped to Phecodes, in subsequent studies, it will be important to further refine these mappings and ensure higher fidelity by establishing new Phecodes where needed. Primary care EHR contains a wealth of information which is not captured in administrative hospital records. For example, Read contains terms on examination findings, laboratory tests, risk factors and symptoms. A potential future expansion of Phecodes could include a custom set of codes to capture these events despite not having a valid ICD-9-CM or ICD-10 target code. This would be relevant not only for UKB participants but also to other deeply phenotyped/genotyped resources e.g. Genomics England (16) or the NHS Digital (17).

Conclusion

In this paper, we introduced our work of mapping the Read2 and CTV3 clinical terminologies used in primary care EHR to Phecodes. We validated our results by replicating previously-reported PheWAS genotype-phenotype associations by performing analyses in the UK Biobank and BioVU. We provide an initial version of the mapping file that can be used by researchers to leverage primary care data for high-throughput translational research.

Acknowledgements

SD, HH are supported by a) Health Data Research UK, b) BigData@Heart, funded by the IMI-2 No. 116074, c) the NIHR UCL Hospitals Biomedical Research Centre. HH, ADH are NIHR Senior Investigators. The dataset(s) used for the analyses described were obtained from Vanderbilt UMC BioVU, supported by institutional funding, private agencies and federal grants incl. NIGMS R01 GM139891, NIA R01AG069900, NIH-funded Shared Instrumentation Grant S10RR025141, and CTSA grants UL1TR002243, UL1TR000445, UL1RR024975. LB is supported by R01-LM010685. Work was supported by the BHF Accelerator Award AA/18/6/24223. ADH supported by Rosetrees and Stoneygate Trust.

370
References


Golden opportunities for clinical decision support in an era of team-based healthcare

Paul R. Dexter, MD,1,2 Titus Schleyer, DMD PhD,1,2

1Regenstrief Institute, Inc., Indianapolis, IN; 2Indiana University School of Medicine, Indianapolis, IN

Abstract

Computerized clinical decision support (CDS) will be essential to ensuring the safety and efficiency of new care delivery models, such as the patient-centered medical home. CDS will help empower non-physician team members, coordinate overall team efforts, and facilitate physician oversight. In this article, we discuss common clinical scenarios that could benefit from CDS optimized for team-based healthcare, including (1) low-acuity episodic illness, (2) diagnostic workup of new onset symptoms, (3) chronic care, (4) preventive care, and (5) care coordination. CDS that maximally supports teams may be one of biomedical informatics’ best opportunities to decrease health care costs, improve quality, and increase clinical capacity.

Introduction

Intractable challenges confront healthcare in the United States,1 including unsustainable rising costs,2,4 primary care physician shortages,5 physician burnout,6 and increasing demand for healthcare services.7 Primary care has been described as in crisis.8 Healthcare is too often inadequately standardized, of poor quality, and unsafe.9,10

Some have suggested that health information technology (IT) holds the promise of driving improvements in healthcare.2,11-13 Others have suggested that solutions to healthcare’s challenges reside in the form of multi-disciplinary coordinated primary care, delivered through team-based approaches such as those found in patient-centered medical homes (PCMHs).14,15 The most effective solutions to healthcare’s critical challenges seem likely to arise from the synergistic combination of CDS and team-based approaches.16 Health IT’s (and CDS’) primary contribution may be in making new care delivery models possible.17

Clinical decision support for multi-disciplinary teams

Excellent opportunities exist for transforming primary care practice by fully leveraging the expertise of clinical teams. Both physicians and non-physician clinicians frequently practice below their level of licensure, with resultant underutilization of nurses and medical assistants.18 Pelak and colleagues estimate that over half of overall care performed by primary care physicians could be safely performed by nurses and other non-physicians.19 Similarly, Altschuler and colleagues estimate that 77% of time spent on preventive care and 47% of time spent on chronic care could be delegated to non-physicians, such as registered nurses, pharmacists, health educators, and medical assistants.20

In the sections below, we describe common clinical scenarios and forms of computerized clinical decision support that could coordinate the parallel efforts of multi-disciplinary primary care clinical teams. We have categorized these primary care tasks into the following clinical domains: low acuity episodic illness, diagnostic work-up of new onset symptoms, chronic care, preventive care, care coordination and other tasks.

Primary care tasks performed by teams utilizing CDS

Clinical domains that could benefit from CDS optimized to support teams include:

Low-acuity episodic illness. As demonstrated by care delivered at an increasing number of retail clinics, nurse practitioners or physician assistants employing guideline-based protocols can effectively manage common low-acuity episodic illness, such as sinusitis and pharyngitis.21,22 Compared with primary care clinics, there is evidence that retail clinics can deliver this healthcare at lower cost with equivalent quality.21,22 In one study, ten conditions commonly treated at retail clinics accounted for an estimated 18 percent of all primary care physician visits.23
Primary care clinics could readily adopt similar guideline-based methods to manage low-acuity episodic illness using computer-based standing orders protocols administered by nurses. An advantage that primary care clinics would have over retail clinics is improved capacity to evaluate persistent symptoms, which in some cases represent serious underlying disease. Computer logic could route these persistent cases to the physician if a scripted telephone follow-up administered by a medical assistant indicated that the symptoms did not resolve or respond as expected.

**Diagnostic workup of new onset symptoms.** For patients who present with symptoms suggestive of more serious underlying disease (e.g., hemoptysis), computer-based scripted methods might provide valuable first-pass diagnostic workup immediately prior to physician evaluation. Through computer-based standing order protocols, preliminary diagnostic tests could be performed prior to the physician's interview of the patient - e.g., a chest x-ray for hemoptysis in an older smoker who has not previously had chest imaging.

Using structured questionnaire data acquired by non-physicians (including from the patient directly) that relate to the patient's symptoms, computer logic could generate a draft differential diagnosis and draft treatment plan for physician review available as the physician goes into the exam room. Such a draft differential diagnosis could list relevant physical exam findings or diagnostic testing that could further narrow the range of diagnostic possibilities.

The availability of first-pass diagnostic approaches acquired through non-physician personnel has the potential to improve upon the estimated 10 to 15% diagnostic error rate. Such efforts might also finally provide the means to insert differential diagnosis generators, such as DXplain, into routine clinical workflow. In turn, draft treatment plans generated by questionnaire data could incorporate recommendations that reflect the best and latest medical evidence, potentially reducing the long delays between definitive research findings and widespread practice. In all cases, it would be the physician or other treating provider who would determine based on their exam if the draft differential diagnosis or treatment plan were appropriate for that particular patient.

**Chronic care.** Guided by electronic questionnaires, non-physicians could capture a much larger portion of history-taking and documentation related to chronic care than currently performed. For example, non-physicians could collect longitudinal patient data necessary for physician decision-making related to asthma therapy, such as the frequency of inhaler use, coughing, or episodes of shortness of breath. They could also capture structured data needed for chronic care quality measures.

To the significant extent that questions are predictable on the basis of a patient's presenting symptoms, the goal would be to use computer logic to collect the answers to those questions directly from the patient (e.g., using tablet technologies) or through a medical assistant. For example, if a patient reports chest discomfort, questions regarding its location, severity, radiation, and association with shortness of breath could follow. Such capture of patient history could relieve the physician of significant portions of visit documentation, allowing the physician to instead focus on verifying the "pertinent positives" from the captured history, supplementing with any needed clarifying questions, and establishing a therapeutic plan with the patient.

For patients with problems identified by a physician, computer logic might also help guide the selection or titration of various medications related to chronic care conditions. An early example of CDS for medication selection in diabetes has been developed by Tarumi et al. Such computer logic has been demonstrated as beneficial for anticoagulants and insulin compared with routine care. There is also demonstrated utility for computer-based hypertension management. Logic could generate medication recommendations to be approved by the physician, or eventually with the proper safeguards, computer-based standing order protocols could directly guide medication titration.

With safeguards in place, a physician might identify a patient in the future with a new diagnosis of hypertension or in need of improved glucose control. That physician could then choose to turn over the choice of medication and/or the subsequent medication titration to non-physicians using computer-based standing order protocols. For purposes of choosing the “right” anti-hypertensive agent, the computer logic itself could search for evidence that the patient has a history of systolic heart failure, diabetic nephropathy, or recent myocardial infarction.

**Preventive care.** Through computer-based standing order protocols, non-physicians could directly coordinate much more preventive care than commonly occurs. In the case of vaccinations, we have previously demonstrated that automated standing order protocols are more effective than physician-directed computer reminders.
Using these approaches, the default for a particular clinic might be that age-specific preventive care is offered to all patients, unless and until such time that the patient in coordination with their physician chooses to stop (e.g., choosing not to proceed with screening colonoscopy due to terminal illness). In cases where the patient refuses recommended preventive care, the reason could be captured electronically and routed automatically to the physician, on the possibility that additional patient education would be useful.

Rather than spending as much time directly ordering preventive care for individual patients, physicians could instead focus on practice-level oversight. There is much potential for saving physician time through such methods given estimates that 7.4 clinician hours per working day are needed to fully follow national recommendations for an average-size panel of 2,500 patients.32

Care coordination and other tasks. Compared to the case where a single clinician is responsible for all orders and documentation, team-based primary care does complicate issues of care coordination. Consequently, the potential value of CDS software is arguably much greater in a team-based environment than in the simpler single physician case. The goal would be that regardless of how work is dynamically allocated or delegated within a practice, or regardless of the sequence by which tasks are completed, there would be software verification that all aspects of healthcare are addressed by the end of the visit.33

Apart from the clinical scenarios described above, many other activities occur frequently in primary care clinics34 that could benefit from team-based CDS software:

- Establishing the patient’s agenda and priorities for the imminent physician visit
- Medication reconciliation
- Reviewing potential side effects of newly prescribed medications with the patient
- Follow-up of hospitalizations and emergency department visits employing standardized questionnaires with forwarding of summaries to the physician.
- Renewals of non-narcotic medications using electronic standing order protocols
- Triage of returning test results, incoming emails and phone calls, insurance clarifications.
- Tracking of referrals from the original order to returning recommendations from the specialist
- Follow-up of missed clinic appointments or diagnostic testing

Team-based approaches hold the key to allowing physicians to spend more of their time overseeing the delivery of healthcare within their practice, rather than personally delivering the vast majority of healthcare. To fully take advantage of the opportunity for physician supervision of practice-level information, much would need to be learned regarding how to optimally summarize the results of team efforts. Valuable reports might include both summary and patient-level data on (1) preventive care administered or refused, (2) blood pressures before and after changes in treatment, (3) medication titrations, and (4) the frequency and nature of various concerning patient symptoms.

Discussion

In team-based care models, CDS software has the potential to be invaluable, transforming potentially fragmented team efforts into highly effective integrated healthcare delivery. Despite risks of fragmented care, the clinical advantages of multi-disciplinary medical care will likely drive wide adoption of team-based practice. CDS that maximally supports teams may be one of biomedical informatics’ best opportunities to decrease healthcare costs, improve quality, and increase clinical capacity.

In this paper, we have outlined clinical tasks that a diverse group of clinicians might effectively address using CDS. Collectively, low-acuity episodic illness, diagnostic workup, chronic care, preventive care, and care coordination account for a large portion of medical practice. Many issues would need to be worked out. For example, full adoption of team-based care approaches might require changes to reimbursement methods, rewarding practices for overall care delivered rather than who delivered what care.

Practices also differ in their precise assortment of team members, so CDS software would need to be highly configurable in terms of which team members are assigned particular tasks. Depending on the practice, available non-physician team members might include nurses, physician assistants, pharmacists, medical technicians,
rehabilitation specialists, social workers, check-in personnel, or patient navigators. In a number of cases, more research would be required prior to wide deployment, such as ensuring that adjustment of anti-hypertensive or diabetic medications employing computer-based standing orders is safe and effective.

Our suggested approaches take advantage of the fact that a significant portion of healthcare is algorithmic in nature, even for the preponderance of cases where “the art of medicine” must subsequently be applied. In a handful of cases, some interventions and clinical scenarios are sufficiently straightforward that they can be safely administered by non-physician clinicians with only high-level provider oversight (e.g., flu shots). For most clinical scenarios, however, the goal would be to simply provide the clinician succinct diagnostic and therapeutic considerations that they can review at the point of care. Ready availability of such considerations at the point of care would best be complemented by trivial methods to directly order those same diagnostic tests or therapies.

The shortcomings of EHR support for team-based care have been previously recognized. The American Medical Association has called for EHR software that (1) facilitates clinical staff to perform work to the extent their licensure and privileges permit, and (2) allows physicians to dynamically allocate and delegate work to appropriate members of the care team. Bates et al. recommends the development of EHR functionality to enable real-time communication and coordination among team members. O'Malley et al. suggests that EHRs could facilitate primary care teamwork by enhancing communication within the practice team, supporting task delegation, and integrating standing orders and protocols. In this paper, we identify specific primary care tasks that could be performed by non-physicians supported by software. We believe these specific examples strengthen the already-compelling case for improving team support in CDS software.

Team-based approaches have been associated with increased physician satisfaction and decreased burnout. Too many PCPs practice within a “frantic bubble,” experiencing workdays as a non-stop stream of patients. Almost all aspects of a primary care physician's busy day could benefit from team-based assistance. CDS software can make that possible.

Acknowledgements: This work was supported by funding from the NIH-NHGRI grant #U01 HG010245, the Indiana Clinical and Translational Sciences Institute (funded in part by Award Number UL1TR002529 from the National Institutes of Health, National Center for Advancing Translational Sciences) Clinical and Translational Sciences Award, and the Lilly Endowment, Inc. Physician Scientist Initiative. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors, and do not necessarily reflect the views of the funding agencies.

References

33. Leveraging all members of the clinical team to improve EHR usability. HIMSS News, 2017.
376


Impact of Clinical and Genomic Factors on COVID-19 Disease Severity

Sanjoy Dey, Ph.D.1,*, Aritra Bose, Ph.D.2,*, Subrata Saha, Ph. D.3,§, Prithwish Chakraborty, Ph.D.1, Mohamed Ghalwash, Ph.D.1, Aldo Guzmán-Sáenz, Ph.D.2, Filippo Utro, Ph.D.2, Kenney Ng, Ph.D.1, Jianying Hu, Ph.D.1, Laxmi Parida, Ph.D.2, Daby Sow, Ph.D.1
1 Center for Computational Health, IBM Research, Yorktown Heights, NY, USA;
2 Computational Genomics, IBM Research, Yorktown Heights, NY, USA;
3 Columbia University Irving Medical Center, Columbia University, NY, USA

Abstract To date, there have been 180 million confirmed cases of COVID-19, with more than 3.8 million deaths, reported to WHO worldwide. In this paper we address the problem of understanding the host genome’s influence, in concert with clinical variables, on the severity of COVID-19 manifestation in the patient. Leveraging positive-unlabeled machine learning algorithms coupled with RubricOE, a state-of-the-art genomic analysis framework, on UK BioBank data we extract novel insights on the complex interplay. The algorithm is also sensitive enough to detect the changing influence of the emergent B.1.1.7 SARS-CoV-2 (alpha) variant on disease severity, and, changing treatment protocols. The genomic component also implicates biological pathways that can help in understanding the disease etiology. Our work demonstrates that it is possible to build a robust and sensitive model despite significant bias, noise and incompleteness in both clinical and genomic data by a careful interleaving of clinical and genomic methodologies.

Introduction

COVID-19 (Corona Virus Disease of 2019), caused by the SARS-CoV-2 virus, is one of the worst pandemics in human history. As of July 6 2021, it has infected 183,700,343 people while claiming 3,981,75 deaths worldwide. The disease exhibits a wide range of symptoms ranging from asymptomatic to mild to rapid progression to critical stage. In general, patients with mild symptoms recover easily whereas severe COVID-19 patients require further treatments such as ventilation in ICU. Such heterogeneity in terms of symptoms poses great challenges for understanding the disease etiology and designing treatment protocols. Characterizing the factors that impact the severity of this illness has been an open research question that we tackle in this work. Typically, severe effects of infectious diseases like COVID-19 are hypothesized to be associated with the variations of host genome. GWAS (Genome wide association studies) have been reported in literature to understand the impact of host genomics on the disease. In addition to viral load environmental factors such as demographics, socio-economic status, lifestyle may also have an impact on the spread, exposure, and severity. The availability of rich Electronic Health Records (EHR) provides an opportunity for picking apart clinical factors associated with COVID-19.

Most of the existing studies that aimed at finding the common risk factors for severe COVID-19 patients focused on either the clinical or the genomics factors, but not on both. In this study, we aim to study both these factors in concert using large-scale datasets that have both EHR and patient genomic data.

Pursuing such joint analysis to find clinical and genomic factors associated with COVID-19 that can be used as biomarkers comes with a few challenges. First, defining COVID-19 severity from EHR is often challenging, as it is not directly collected in these datasets. Hence, the severity has to be inferred using appropriate clinical knowledge as a surrogate phenotype which in turn may introduce label noise and data collection bias. Thus, building machine learning models on clinico-genomic dataset needs special adjustments to handle such noise and bias associated with COVID-19 severity. Second, assessing the impact of genomic factors on COVID-19 severity may be confounded by several other clinical factors such as the prior comorbidities of patients and treatment protocols. The effect of such confounding factors has to be addressed carefully when conducting the combined clinico-genomic studies. Third, the impact of clinico-genomic factors may also vary depending on the variations of diverse COVID-19 strains that have been observed due to genetic mutations.

* Lead authors with equal contributions.
§ Author contributed while he was at IBM Research, Yorktown Heights, NY.
We propose in this work a combined framework for finding the clinical and genomic factors associated with COVID-19 severity using a large COVID-19 dataset from the UK Biobank (UKBB). To address the above mentioned challenges, we first use the COVID-19 related hospitalization as a surrogate outcome for defining severe COVID-19 cases. Second, we use a machine learning technique called positive-unlabeled (PU) learning to address the noise and reporting bias present in COVID-19 severity labels. Moreover, we use a recently proposed genomic analysis framework entitled RubricOE to select the set of genomic factors after adjusting for the common prior comorbid conditions which may act as potential confounders. Finally, we aim to assess how the importance of the extracted biomarkers evolve over the pandemic marked by events such as the introduction of using Dexamethazone as treatment for COVID-19 severity and the emergence of the reportedly more contagious B.1.1.7 SARS-CoV-2 (alpha) strain.

Related Works

To combat the global pandemic caused by COVID-19, scientists across different disciplines have been studying the disease to understand various facets such as its spread, epidemiological and patho-physiological characteristics, and societal impact.

To help consolidate and streamline the efforts around studying the host genetics for susceptibility, the COVID-19 Host Genetics Initiative was launched in mid 2020 with a primary focus on disease severity. As part of this initiative, Hou et al. found that ACE2 or TMPRSS2 DNA polymorphisms were likely associated with genetic susceptibility of COVID-19, while Li et al. found that genetic variants under a polygenic model shows promising improvements in prediction accuracies. Wang et al. presented further genetic insights into phenotypic difference among the COVID-19 patient groups, highlighting genes and variants of interest. More specifically, they conducted both single-variant and gene-based association with respect to five severity groups finding insights such as variants involved in IL-1 signaling pathways. While studies identified a gene cluster on chromosome 3 as a risk locus for respiratory failure after infection with COVID-19, Zeberg et al. found that the risk is conferred by a genomic segment that is carried by around 50% of people in South Asia and 16% of people in Europe.

Simultaneously, there have been concerted efforts to characterize the clinical characteristics of COVID-19 including the patho-physiology and the risk factors for COVID-19. Several large-scale meta-analysis studies were conducted either on previous studies or by summarizing published articles to identify common clinical symptoms and lab abnormalities among severe COVID-19 patients. Some common risk factors observed among critical/mortal COVID-19 patients are socio-demographic information such as male, older than 65, smoking, and prior history of hypertension, diabetes, cardiovascular disease, respiratory diseases, kidney diseases, obesity, immunosuppresion, and cancer. In addition to these prior conditions, depression as well as other cognitive and neurological disorders were also found as risk factors by a nationwide cohort study in Israel, although this study found that smoking and presence of respiratory diseases do not significantly increase the risk of complications. These factors were also found in a separate study performed on 40% of all patients in England. In addition, socio-economic deprivation, diabetes, severe asthma, Black and South Asian ethnicity were observed as high risk factors. Research has also been conducted to characterize risk among sub-populations such as obesity, heart failure, and Parkinson’s Disease concurrently. Mathew et al. found three immunotypes revealing different patterns of lymphocyte responses among hospitalized COVID-19 patients.

Besides finding either clinical or genomic risk factors of severe COVID-19, several machine learning (ML) and artificial intelligence (AI) have also been applied successfully to solve a wide range of needs, ranging from fast ML based COVID-19 infection prediction to routine blood-test as fast alternatives to costly and time-consuming PCR tests to predicting disease state of individual cells given their transcriptome. However, to the best of our knowledge, we did not find any machine learning model built on combined clinico-genomic dataset to assess the COVID-19 severity.

Method

Dataset. We analyzed the large prospective cohort of patients from the UK collected in the UKBB repository. The dataset contains diverse information including demographics, diagnosis, medications, lab tests, and genomic information of approximately half a million patients. In addition, the national SARS-CoV-2 laboratory test data were made available in UKBB through the Public Health England (PHE). This COVID-19 dataset contains a flag indicating specimen origin of COVID-19 tests: hospital inpatient origin vs other settings. It also notes the specimen collection.
Table 1: The distribution of unique patients based on Specimen Origin and result as of Jan 21, 2021

<table>
<thead>
<tr>
<th>Specimen Origin</th>
<th>Result=0</th>
<th>Result=1</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Specimen Origin=0</td>
<td>11,989</td>
<td>9,083</td>
<td>21,072</td>
</tr>
<tr>
<td>Specimen Origin=1</td>
<td>31,800</td>
<td>4,345</td>
<td>36,145</td>
</tr>
<tr>
<td>All</td>
<td>43,789</td>
<td>13,428</td>
<td>57,217</td>
</tr>
</tbody>
</table>

date and the *specimen result* in addition to a few other information about how the specimen was collected. Data were available from Mar 16, 2020 and the repository has been updated once or twice every month. In our current study, we used data from Mar 16, 2020 to Jan 21, 2021.

**Preparing the outcome and feature sets.** Although the COVID-19 dataset in UKBB is a prospective study, it did not collect the disease severity explicitly. The nature of this dataset is very diverse depending on national testing strategy which evolved over time as the pandemic progressed. For example, in UK, testing was initially restricted to those with symptoms in hospital and thus under this assumption, using the positive tests of these subjects in hospital can be a reasonable proxy for severity. However, when the testings were expanded covering even asymptomatic patients from diverse facilities, defining the severity required analyzing information regarding the specimen sources of test.

Among several different fields available in the COVID-19 table of our the UKBB, we settled on *specimen origin* and *specimen results* to track severity. We have used a combination of these two flags for assessing disease severity under the assumption that patients with a positive COVID-19 test obtained in hospital (i.e., *specimen origin=1*) are likely to be severe cases than those COVID-19 patients who were tested in outpatient services (i.e., *specimen origin=0*). Table 1 contains the number of unique patients who had at least one positive test result and at least one ‘inpatient’ indicator retrieved from the origin field. We used inpatient COVID-19 samples (*specimen origin=1 and result =1*) as the severe COVID-19 patients while other COVID-19 patients (*specimen origin=0 and result = 1*) as the non-severe cases.

While this approach estimates severity, it may be prone to a selection bias problem by defining all inpatient COVID-19 patients as severe cases. For instance, some patients may have been hospitalized due to other conditions or may have been experiencing other problems before they were COVID-19 positive. To cope with aspects of this problem, a separate field covering hospitalization-acquired infections was collected in the database to filter out such patients from actual severe COVID-19 infection related hospitalization. However, this field was enabled in the dataset very recently and the protocol for marking that field is not yet uniformly applied to all samples.

Another potential issue with the dataset pertains to defining the non-severe COVID-19 patient cohort. While hospitalization data for acute COVID-19 patients were collected carefully, the non-severe patients are harder to define. For example, patients who tested positively for COVID-19 in outpatients or non-urgent care service centers (*specimen origin=0*) may have experienced severe symptoms. The lone presence of the Origin flag set to 0 does not always mean that patients did not have severe reaction at a later stage, in fact, they may not have been traced prospectively in the dataset. Such tracing would most likely require conducting another test while being in an inpatient service. To cope with this issue, we used a special class of machine learning to handle this kind of noise associated in the outcome labels.

To build our severity prediction models, we used several clinical and genomic features that may have an impact on COVID-19 severity. For the clinical features, we manually curated some of the most relevant features which may have association with COVID-19 for further analysis. Our clinical features include demographics, lifestyles, comorbidities based on patient’s prior history, self-reports, and their chronic disease histories curated from prior inpatient records.

**RubricOE: the Omic Framework.** The pipeline of RubricOE (learning rubric for multi-omics Genetic Epidemiology) is outlined in Figure 1. We describe data treatment in two stages as follows.

1. **Pruning Genomic Features:** Starting with 12,965 overlapping samples with genomic information in the UKBB COVID-19 dataset, we only retained high quality imputed genomic markers in the form of Single Nucleotide Polymorphism (SNPs) with imputation quality INFO score > 0.3, resulting in approximately 4.4 million SNPs. We further
did a standard quality control (QC) for filtering informative variants, removing SNPs as well as individuals with more than 2% of missing data. We further filtered for Minor Allele Count (MAC) > 0.05, sex discrepancy in individuals, variants deviating by more than $p < 1e^{-6}$ in the Hardy-Weinberg equilibrium (HWE) for cases and $p < 1e^{-10}$ in controls. We also filtered for individuals with ±3 standard deviation in heterozygosity rates and very closely related individuals (second degree). All of the QC analysis was done using PLINK v1.9. We ended up with 12,389 individuals and 4,539,795 SNPs which passed all the QC thresholds. Of these 12,389 individuals, 4,000 were cases with severe COVID-19 infection and 8,389 were non-severe individuals who tested positive. We further pruned for linkage disequilibrium (LD), removing correlated SNPs with $r^2 > 0.2$ with a variance inflation factor of 10 on a window size of 50 kb. We performed a Genome-wide Association Study (GWAS) on this pruned dataset for a sanity check of the QC protocol and observe significant loci associated with the severity of COVID-19 infection. GWAS was performed by the package SAIGE, while correcting for population structure, sex and age. We computed the top 20 principal components (PCs) with TeraPCA as a proxy for population structure and used them as covariates along with the biomarkers.

2. **Stable Feature Extraction:** The learning rubric employs a nested test-training set configuration on the genomic data after QC and pruning for LD. The outer test-training set split reserves the test set (“validation”) for final SNP evaluation and denotes the training set as “working” data. Within the latter, further train-test splits are performed to rank the features by their Youden index (also known as J statistic). RubricOE is applied on the genomic data, accounting for top 20 PCs, sex, year of birth and Diabetes Mellitus (DM) as covariates. It iteratively finds a “stable” set of SNPs which persistently remains highly ranked with their respective Youden index. Annotation, Gene Ontology (GO) and enriched pathway analysis of the “stable” set of SNPs produced by RubricOE has been evaluated using the R package clusterProfiler.

**Clinico-Genomic Framework.** We used a predictive learning framework called Positive-unlabeled (PU) learning for extracting the most significant clinical and genomic features associated with COVID-19 outcome. Note that our dataset is more robust in terms of defining the severe COVID-19 patients than the non-severe cases. Hence, using a classical binary classification algorithm will not be appropriate and able to cope with such noisy class labels. PU learning algorithms have alternatively been reported to be especially suitable for learning from a noisy negative class by treating instances in this class as unlabeled while predicting the positive class (severe COVID-19 cases).

PU learning is a variant of the classical binary classification problem, where it is assumed that the unlabeled examples could belong to either positive or negative class. We refer to [30] for a comprehensive review of such PU learning.
techniques. Among several state-of-the-art PU learning techniques, we use PU Adapter, uPU, and nnPU for learning the COVID-19 severity. All of these PU learning techniques have the basic assumption that the observed samples are drawn uniformly from the positive distribution, which enables PU learning techniques to leverage standard binary classification methods with minor modification of data or algorithm.

The PU-Adapter technique pre-processes the available PU data so that any standard binary classification method can be used for learning on the PU dataset. In particular, this method reweights the samples of PU data so that learning a traditional binary classifier on the weighted samples yield to similar target probability threshold on the PU dataset. Moreover, it has been shown that the classifier trained on positive and unlabeled examples predicts probabilities that differ by a constant factor from the true probabilities of positive class. Unbiased PU learning (uPU) uses two separate loss functions for the positive and unlabeled samples using convex optimization framework.

Model Evaluation. We evaluate the effectiveness of PU-learning algorithms with a comparison with baseline traditional classification algorithms using logistic regression with a $L_1$ loss. Computing traditional metrics such as precision, recall, $F_1$ score, and accuracy is challenging in the positive unlabeled scenario since the only information available relates to the positive label while no sample from the negative class is clearly provided. One straightforward solution to this problem is to assume that the unlabeled data are negative. However, this is not fully accurate. A modified $F_1$ score (we call it as $F_1^{pu}$ score) has been proposed to address this issue. This score is computed as $r^2 / (y + 1)$, where $r$ is the recall and $p(y = 1)$ is the prevalence of the predicted positive label. As shown in, the $F_1^{pu}$ score can take values greater than one but shares the same property as $F_1$ score; it is high when both precision and recall are high and small when either one is small. We also report metrics used to evaluate traditional classification methods for completeness although these are not recommended for our setting. They include accuracy, precision, recall, and $F_1$ score. We used a 5-fold cross-validation (CV) approach for evaluating the performances of the predictive models, where in each CV fold, the hyper-parameter for $L_1$ loss was tuned internally using a grid-search over the range of $[10^{-3}, 10^{-2}, 10^{-1}, 1, 10]$. In particular, for clinico-genomic feature importance, 100 bootstrapping runs were used to estimate their confidence intervals. We used DPM360 workbench to conduct the analysis in a repeatable and reusable manner.

Results

The results of the baseline $L_1$ regularized logistic regression are shown in Table 2. We report the metrics for three different models: clinical, genomic, and the combined clinico-genomic model. Overall, we can observe that clinical model has worse predictive power than the genomic model. However, the clinico-genomic model shows significant improvement in terms of predictive power.

Table 3 and Table 4 show the same metrics from the three different models using two different PU learning techniques: nnPU and PU-Adapted logistic regression, respectively. The nnPU model does not perform well for categorical genomic features, moreover it is harder to interpret such model. In contrast, the PU-adapted method uses logistic regression as the baseline model. Hence it is easier to interpret the resulting coefficients as log-odds ratios of clinico-genomic features. The combined clinico-genomic model of the PU adapted logistic regression model yields significant improvement from the models built on clinical and genomic model individually. Although the $F_1$ score of the PU learning model is similar to the score of the baseline logistic regression model, the adjusted $F_1$ score is higher for the PU model, which justifies the use of the PU model for our noisy class level.

Figure 2: Top 30 statistically significant features obtained from PU-Adapted Logistic Regression. Plots show mean coefficients (log odds ratio) for regression along with 95% bootstrapped confidence interval.
Table 2: Average (std. deviation) 5-fold CV performances of three models based on Logistic Regression with $L_1$ loss

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>Accuracy PU</th>
<th>$F_1$ score</th>
<th>Accuracy</th>
<th>$F_1^{PU}$ score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinical</td>
<td>0.488 (0.008)</td>
<td>0.629 (0.014)</td>
<td>0.768 (0.006)</td>
<td>0.550 (0.005)</td>
<td>0.667 (0.005)</td>
<td>0.949 (0.019)</td>
</tr>
<tr>
<td>Genomic</td>
<td>0.507 (0.023)</td>
<td>0.665 (0.013)</td>
<td>0.799 (0.018)</td>
<td>0.575 (0.019)</td>
<td>0.684 (0.009)</td>
<td>1.045 (0.035)</td>
</tr>
<tr>
<td>Clinico-Genomic</td>
<td>0.564 (0.023)</td>
<td>0.710 (0.030)</td>
<td>0.875 (0.022)</td>
<td>0.629 (0.026)</td>
<td>0.730 (0.009)</td>
<td>1.241 (0.057)</td>
</tr>
</tbody>
</table>

Table 3: Average (std. deviation) 5-fold CV performances of three models based on nnPU Model

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>Accuracy PU</th>
<th>$F_1$ score</th>
<th>Accuracy</th>
<th>$F_1^{PU}$ score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinical</td>
<td>0.586 (0.017)</td>
<td>0.416 (0.019)</td>
<td>0.803 (0.003)</td>
<td>0.486 (0.009)</td>
<td>0.716 (0.006)</td>
<td>0.754 (0.022)</td>
</tr>
<tr>
<td>Genomic</td>
<td>0.4 (0.545)</td>
<td>0 (0.000)</td>
<td>0.677 (0)</td>
<td>0.001 (0.001)</td>
<td>0.677 (0.000)</td>
<td>0.004 (0)</td>
</tr>
<tr>
<td>Clinico-Genomic</td>
<td>0.830 (0.025)</td>
<td>0.054 (0.010)</td>
<td>0.694 (0.003)</td>
<td>0.100 (0.018)</td>
<td>0.691 (0.003)</td>
<td>0.138 (0.027)</td>
</tr>
</tbody>
</table>

Table 4: Average (std. deviation) 5-fold CV performances of three models based on PU Adapted Logistic Regression

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>Accuracy PU</th>
<th>$F_1$ score</th>
<th>Accuracy</th>
<th>$F_1^{PU}$ score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinical</td>
<td>0.401 (0.008)</td>
<td>0.800 (0.011)</td>
<td>0.621 (0.022)</td>
<td>0.534 (0.006)</td>
<td>0.548 (0.015)</td>
<td>0.991 (0.028)</td>
</tr>
<tr>
<td>Genomic</td>
<td>0.417 (0.004)</td>
<td>0.862 (0.006)</td>
<td>0.660 (0.010)</td>
<td>0.562 (0.005)</td>
<td>0.567 (0.007)</td>
<td>1.114 (0.018)</td>
</tr>
<tr>
<td>Clinico-Genomic</td>
<td>0.485 (0.006)</td>
<td>0.839 (0.021)</td>
<td>0.794 (0.012)</td>
<td>0.615 (0.009)</td>
<td>0.660 (0.007)</td>
<td>1.260 (0.040)</td>
</tr>
</tbody>
</table>

We interpret the log-odds of the PU-Adapted logistic regression model to denote the importance of a particular feature toward COVID-19 severity. Figure 2 shows the feature importance of top 30 statistically significant features associated with COVID-19 severity computed by the final combined clinico-genomic model using the PU learning framework. The statistical significance of a feature is computed using bootstrapping with 100 runs and the confidence interval is represented as error bars in the figure which are beyond zero (the vertical line).

Discussion

Our final combined clinico-genomic model finds the significant features associated with COVID-19 severity (Figure 2), which may act as potential biomarkers. Some of the top significant clinical features have already been found in previous studies. For example, age and black ethnicity have been reported in studies on COVID-19 and prior conditions like diabetes mellitus, hypertension have been reported in previous studies. Note that our study finds smoking status (past smoker, current smoker, pack years of smoking) is related to severe COVID-19 patients, which has ambiguous evidence in literature. A few studies have reported it with increased risk factors, while others did not find it to be a significant risk factors for COVID-19 severity.

In addition, our study found a few more features such as number of medication taken per day to be significant. Such features are related to the overall health status of the patient and correlate indirectly to high severity risk factors. Interestingly, distance from home to workplace has also been reported with lower risk of severity, which may be confounded by people commuting by car for longer distance instead of public transport or cycling for shorter commutes.

Is the clinico-genomic model sensitive to treatment protocols? UKBB does not have direct information on COVID-19 drugs and detailed treatment protocols that a COVID-19 patient went through, so we assess the impact of a COVID-19 drug indirectly using its date of approval. In particular, we study the change in coefficients of the top features (Figure 3) before and after the drug approval date. Dexamethazone was approved for use in UK on Jun 15, 2020. We observe (see Figure 3) that most of the significant factors from the whole dataset still had similar impact after the drug being used, but only 5 features—age, current smokers, DM, gene MACROD2 and SNP rs10733122—had significant impact before the approval date. MACROD2 is known to be associated with ischemic stroke and large-artery atherosclerosis, both of which are correlated with COVID-19 severity. Similarly, age, DM and smoking status are also shown to be affected by corticosteroids including Dexamethazone. While we observe that the model is sensitive to treatment protocols in general, it should be noted that this analysis may be confounded due to small sample size or other environmental factors linked with the use of corticosteroids. Also,
medication data were not directly available in the cohort we used - as such this analysis has a strong assumption that the treatment strategy was majorly adopted for severe COVID-19 patients after it was approved. We aim to explore the effect of medication and treatment procedure more systematically in future.

**Is the clinico-genomic model sensitive to the evolving virus?** UKBB did not collect the virus genome prospectively. So we use the date of diagnosis as a proxy for the dominant strain. In the UK, a reportedly more contagious strain named B.1.1.7 started surfacing from early September 2020 and it is estimated that by December 2020, over 60% of new COVID-19 patients had the newer strain. We divided the whole cohort into three: those diagnosed before Aug 31, 2020 (older strain), between Sep 1st, 2020 and Dec 15, 2020 (both strains), and, after Dec 15, 2020 (newer strain). Figure 4 shows how the effect of the top clinico-genomic features change over these three cohorts. We observe that some factors like age, prior DM condition and number of medication are consistent across both strains, while the smoking history (current and present) had larger impact on older strain than the newer strain. Such analysis of temporal trend of a factor’s impact on severity can help elucidate important clinical insights.

**Biological implications of the clinico-genomic model.** The genomic features obtained from the model are enriched in a host of biological pathways related to transmembrane transporters and receptor activities (Figure 5(a)). The most significant is transmembrane receptor protein tyrosine phosphatase which on increasing activity might affect T cells and contribute to their depletion and immunoparalysis in severe COVID-19 patients. Other important pathways such as ion channel, metal ion and inorganic cation transmembrane transporter activities reaffirm the notion that SARS-CoV-2 E protein is a potential ion channel like other coronaviridae. Focusing on the significant biomarkers from the combined model, we observed two new GO enriched terms, passive transmembrane transporter activity and PDZ domain binding (Figure 5(b)). It has been shown that PDZ-containing proteins among binders of the SARS-CoV-2 proteins E, 3a or N affect viral replication under knock-down gene expression in infected cells, significantly particularly in the E protein. The model highlights genes such as CDC42BPA and LARS2 which are previously shown to be associated with COVID-19 antibodies and respiratory failure caused by COVID-19 respectively. It also highlighted regions in Chromosome 19 which is associated with critical illness caused by COVID-19 and possible candidate chromosomal regions in Chromosomes 15, 8 and 4 which can be linked with COVID-19 severity when taking clinico-genomic factors into account.
Figure 4: Mean coefficients with 95% CI of features from Fig 2 indicating severity of COVID-19 with dominance of (a) older strain, (b) both strains, and (c) newer strain (B.1.1.7) in UK. Features that are still significant (not significant) are shown via deep grey (light grey) bars with diamond (circular) markers.

(a) older strain
(b) both strains
(c) newer strain

Figure 5: Boxplots of significant GO enriched pathways, colored by their corresponding corrected p-value and count of mapped genes in the x-axis.

(a) Pathways of “stable” SNPs from RubricOE.
(b) Pathways of significant SNPs from combined model.

Conclusion

In this study, we extracted significant clinical and genomic factors associated with severe COVID-19 from a large-scale EHR and genomic dataset available from UK. In particular, we used a special class of machine learning methods called positive-unlabeled learning to address the challenge of noisy class label of COVID-19 severity outcome, and paired it with a state-of-art omic analysis tool, RubricOE, to enable efficient use of the limited amount of genomic data available. We demonstrated significant improvement in disease severity prediction using this novel clinico-genomic modeling. We also showed that the model is sensitive to patient treatment, sensitive to the evolving virus and implicates biological pathways that can even potentially inform drug-targeting. Thus collectively the results can be used in discovering potential biomarkers for better clinical decision making. The novel clinico-genomic modeling combination presented in this paper represents a step forward in combining clinical and genomic data towards clinical insights, and has implications beyond COVID-19, as many of the limitations in clinico-genomic data we sought to overcome are common across many disease areas.

Specifically with COVID-19, in the future we plan to further investigate the effect of treatment protocols as even more data at higher resolutions become available including effects of vaccination, COVID-19 induced deaths and long-COVID effects.

References


38. Tomar PPS, Arkin IT. SARS-CoV-2 E protein is a potential ion channel that can be inhibited by Gliclazide and Memantine. Biochemical and Biophysical Research Communications. 2020;530(1):10–14.
Developing a Data Quality Standard Primer for Cardiovascular Risk Assessment from Electronic Health Record Data Using the DataGauge Process

Franck Diaz-Garelli, Ph.D.,1 Andrew Long, M.P.H.,1 Michael P. Bancks, Ph.D. M.P.H.,2 Alain G. Bertoni, M.D.,2 Adhithya Narayanan, B.S.,3 Brian J. Wells, M.D. Ph.D.2
1University of North Carolina at Charlotte. Charlotte, NC
2Wake Forest School of Medicine, Winston Salem, NC
3Columbia University, New York, NY

Abstract

The learning health systems aim to support the needs of patients with chronic diseases, which require methods that account for electronic health recorded (EHR) data limitations. EHR data is often used to calculate cardiovascular risk scores. However, it is unclear whether EHR data presents high enough quality to provide accurate estimates. Still, there is currently no open standard available to assess data quality for such applications. We applied the DataGauge process to develop a data quality standard based on expert clinical, analytical and informatics knowledge by conducting four interviews and one focus group that produced 61 individual data quality requirements. These requirements covered all standard data quality dimensions and uncovered 705 quality issues in EHR data for 456 patients. These requirements will be expanded and further validated in future work. Our work initiates the development of open and explicit data quality standards for specific secondary uses of clinical data.

Introduction

The development of learning health systems depends on reliable secondary use of Electronic Health Record (EHR) data.1,2 Research endeavors including comparative effectiveness research,3,4 precision medicine5–7 as well as the delivery of precision care8,9 relies on EHR secondary use. Populations at risk for chronic diseases are among those that would benefit most from such approaches via clinical data reprocessing to provide further support to clinicians9,10 For example, patients with diabetes (12–14% of U.S. adults)11 are at increased cardiovascular disease (CVD) risk12,13 and likely to benefit from routine personalized CVD risk prediction to guide clinical care due to the changing nature of their risk factors over time.14 In the context of learning health systems, this can be explored efficiently by leveraging existing EHR databases.1,2

EHR-based risk assessment approaches hold great promise to improve care and outcomes via patient-centric learning health care practices including data-driven personalized medicine,5–7 real-time clinical decision support15 and real-world evidence-based comparative effectiveness-base practices.4 However, EHR data quality remains a core barrier to this goal.1,3,16 Data quality issues such as the pervasiveness of incomplete data is a direct threat to risk prediction reliability.4,17–19 Also, discordant data sources within and across EHRs can generate uncertainty on patient information,20,21 which reduces confidence in EHR-extracted conclusions. It has been shown that there are significant differences between independent research cohort data and EHR data quality,22 which raises further questions for secondary uses. At present, it is known that data quality has the potential to impact risk prediction accuracy23,24 but the extent is yet to be quantified and accounted for to ensure accurate risk assessment.16 A major barrier to achieving this is the lack of an explicit, validated data quality standard to assess EHR data destined for CVD risk assessment.

In this paper, we present our initial evaluation results and data quality assessment standard prototype for the assessment of EHR data destined for CVD risk assessment. We built this standard using the DataGauge process25 to design and implement data quality assessments for EHR data. This method defines an explicit data quality standard including a data model paired to discrete data quality requirements based on qualitative research methods and expert knowledge that enable open and transparent reporting of data quality evaluation parameters and data quality assessment results.26 We developed our standard for the Framingham 10-year hard coronary heart disease risk prediction model27 based on expert interviews and a focus group session. Our standard was validated for coverage of existing data quality framework features including data quality dimensions, levels of data granularity and knowledge domains necessary for data quality assessment.26,28–30 This standard represents the first step toward open, standardized data quality assessment of EHR data destined to estimate CVD risk in the context of learning health systems. Our aim is to pave the way toward open development and evaluation of data quality assessments for secondary analyses of clinical data in learning health systems.
Methods

To develop our data quality gold standard, we followed the DataGauge process.\textsuperscript{25} We selected this approach because it provides a systematic methodology to generate an explicit data quality standard, ensuring transparency\textsuperscript{26,31} in reporting. This approach also ensures coverage of a broad set of aspects required for exhaustive and complete data quality assessment.\textsuperscript{25,29,30,32} Applying the DataGauge process entails iteratively developing a model to explicitly frame the evaluation scope and clearly define the data needs for a specific secondary use. In our case, our intended use was leveraging EHR data to calculate patients’ CVD risk using the Framingham 10-year hard coronary heart disease event risk model.\textsuperscript{27} This risk assessment model was selected for its coverage of many factors used for other risk scores such as the Atherosclerosis and Cardiovascular Disease Pooled Cohort Equations\textsuperscript{33} and the Multi-Ethnic Study of Atherosclerosis (MESA) risk score.\textsuperscript{34} The model was also selected for its relative simplicity compared to other models, which fit the purpose of this initial pilot study.

The DataGauge process entails developing an explicit list of data quality requirements using qualitative research methods based on the initial data needs model. We developed these requirements using a two-stage approach involving interviews and focus group sessions.\textsuperscript{35} Specifically, we developed an initial interview guide based on the third version of the data needs model and the guidance provided with the DataGauge methodology along with three rounds of feedback from informaticians and qualitative researchers. This aimed to ensure the coverage of data quality dimensions and levels of data granularity.\textsuperscript{25,30,36} To ensure coverage of all areas of expertise needed for this secondary analysis of EHR data, we recruited one informatician, one epidemiologist, one clinician and an interdisciplinary expert covering all three areas. Following these four initial interviews, we developed a focus group guide with three rounds of review. We then conducted a focus group session including all four experts. The goal of this session was to ensure agreement across data quality requirements and discuss issues across areas of expertise. Our qualitative research approach was based on grounded theory.\textsuperscript{35} We employed a purposive sampling approach where experts were selected based on their areas of expertise but also were familiar with secondary analyses of clinical data, in accordance with the DataGauge process guidance. Saturation was ensured by providing relatively ample time for interview sessions (one hour each), multiple passes on the interview guide and allowing for additional expert review of the data quality requirements to ensure that no additional requirements could be generated.

We evaluated the resulting data quality standard in two ways. First, we evaluated the coverage of our data quality requirements based on existing EHR data quality assessment frameworks. Specifically, we assessed our requirements’ coverage of data quality dimensions,\textsuperscript{25,29,32} levels of data granularity\textsuperscript{25,36,37} and knowledge domains required for secondary analyses of clinical data (i.e., clinical, analytical and informatics), in agreement with the DataGauge methodology’s framework.\textsuperscript{25,30,38} As an external validation, we also assessed the coverage of dimensions reported by Weiskopf’s 3x3 DQa framework\textsuperscript{28} which includes completeness, correctness and currency as data quality dimensions and patient, variable and time as data constructs to be covered for any data quality assessment. Coverage was represented by the number of areas touched on by the requirements. The requirements were classified along each dimension by two raters and then we calculated inter-rater agreement\textsuperscript{39} and two-rater Cohen’s Kappa\textsuperscript{40} to verify inter-rater reliability across methods using the irr R package.\textsuperscript{41} Discrepancies were discussed and reconciled for the final report. Then, we conducted an initial effectiveness evaluation by implementing the data quality assessment as defined by our data needs model and quality requirements. The goal was to test whether our data quality standard was capable of catching data quality issues in real-world EHR data. Specifically, we extracted all EHR data elements needed for CVD risk scoring for 456 patients. We included all data for all patients between 45 and 54 years of age, with ten or more hypertension diagnoses, from the Wake Forest Baptist Health system (Winston-Salem, NC) for this pilot evaluation. The age range was selected to fall well within the target age range of the CVD risk assessment model. Hypertension was selected due to its known indirect interaction with CVD\textsuperscript{26} as well as the need for long-term follow-up; the ten-diagnosis cutoff was selected to ensure that the patient was followed in the EHR for a minimum of 10 visits and would contain enough data to assess data quality but also to limit the number of patients in the evaluation dataset. The data was reshaped to conform to the data needs model and evaluated according to the data quality requirements using custom-built code including queries developed.

Multiple software tools were used to carry out this analysis. Data extraction was done using Wake Forest Baptist Health’s 12h\textsuperscript{41} clinical data warehouse. The data needs model visualizations were created using MySQL Workbench (version 8.0.23, Oracle, Corp., Austin, TX). The same tool was used to shape data extraction to fit our analytical data model. Rater analyses were done in R version 3.6.2\textsuperscript{40} and RStudio (version 1.2.5033, RStudio, Inc., Boston, MA). Visual exploration and analyses were performed using Tableau (version 2020.4.2, Tableau Software, Inc., Seattle,
Results

Our final data needs model was established after three rounds of feedback (Figure 1). The final model was a single- table design representing a patient data “snapshot” containing the latest data available for the calculation of a CVD risk score at the time of a visit. This design was chosen to align with the data’s intended purpose and also as a means to facilitate the assessment of the data’s fitness for purpose.\(^\text{31,45}\) Multiple values can be available over time or on the same visit despite needing a single value for CVD risk assessment and our data model addresses this data management concern to simplify analysis and quality assessment. The variables included are a direct reflection of those for CVD risk calculation as defined by the Framingham 10-year hard coronary heart disease event risk model.\(^\text{27}\) This data needs model was subsequently used to develop data quality requirements.

Developing our data quality requirements in accordance with the DataGauge process entailed conducting four interviews and one focus group session. The interview lengths ranged between 49 and 87 minutes and were conducted between 01/19/21 and 02/10/21. Our focus group lasted 59 minutes and was conducted on 02/18/21. The interviews produced 47 individual data quality requirements. The focus group generated 14 new requirements and allowed the adjustment and fine-tuning of 12 requirements. Table 1 presents the list of all expert feedback leading to primary quality requirements along with the exhaustive list of data quality requirements for EHR data destined to calculate patients’ CVD risk over time using the Framingham 10-year hard coronary heart disease event risk model.\(^\text{27}\) The final list of requirements contained value-level requirements such as type checks, value checks, range checks and NULL value checks (e.g., requirement ID 1, 2, 6, 10 and 49) as well as more complex requirements such as normality checks at the variable level (e.g., requirement ID 4), cross-variable checks such as (e.g., requirement ID 38) and patient-level checks for alignment with clinical knowledge (e.g., requirement ID 41).

Our evaluation results confirmed that the requirements generated using the DataGauge process were capable of covering all targeted aspects of data quality. The classification of our requirements by two reviewers yielded inter-rater agreement and reliability of 95.1% (Kappa=0.977, p<.0001), 90.2% (Kappa=0.879, p<.0001) and 100% (Kappa=1, p<.0001) for data quality dimensions, knowledge domains and levels of data granularity, respectively. Classification of the 3x3 DQA constructs was slightly more challenging but feasible with reasonable levels of inter-rater agreement, returning agreement values of 88.5% (Kappa=0.801, p<.0001) and 90.2% (Kappa=0.924, p<.0001) for data quality dimension and construct, respectively. Figure 2 presents the coverage of data quality dimensions, knowledge domains necessary for secondary analyses of clinical data and levels of data granularity from the DataGauge Framework but also the coverage of all construct combinations from the 3x3 DQA framework, as an external validation. Coverage ranged from 2 to 26 requirements across data quality dimensions, from 2 to 18 for knowledge domains and from 5 to 36 for variable granularity level. These variations are expected and due to the particular aspects of the data quality assessment design. Still, all aspects were covered by at least 2 requirements. Our requirements also covered most combinations of the 3x3 DQA framework. Coverage ranged from 22 requirements for variable-level requirements checking for correctness to 1 patient-level requirement checking for currency. The process did not generate any variable-level requirement checking for data currency, likely due to the data needs model design that split variables and their corresponding temporal variables, allowing for variable-level definitions of accuracy to cover this.

On implementing our data quality requirements to assess real EHR data, we found that our requirements were able to identify data quality issues that would impact or prevent the calculation of CVD risk assessment. Our dataset contained data for 456 patients followed for at least one year with a mean (± standard deviation) age of 57 (±16) years. The unprocessed dataset contained 506,268 diagnosis entries (4560, relevant to our quality assessment), 732,925 vital value entries (213,294, relevant), 3,194,129 medication entries (18,765, relevant) and 1,178,552 lab value entries (6786 relevant). The data quality assessment revealed that our requirements were capable of uncovering 705 data quality issues in the data representing this patient set. These issues included 4 data type violations (e.g., requirement ID 1), 350 NULL value violations (e.g., requirement ID 53), 5 value range violations (e.g., requirement ID 24), 12 patient ages outside of usable range (e.g., requirement ID 33), 332 timing of values harvested (e.g., requirement ID 36) and distribution issues for 2 variables (e.g., requirement ID 46).
Figure 1 – Data Needs Model Versions. Figure (c) presents the final data needs model as used to develop the data quality requirements list.

Figure 2 – Coverage of Data Quality Dimensions, Data Reuse Knowledge Domains, Data Granularity Levels and Construct combinations from the 3x3 DQa framework.
<table>
<thead>
<tr>
<th>ID</th>
<th>Variable</th>
<th>Data Quality Requirement</th>
<th>Expert Reviewer Quote (Interview &amp; Focus Group)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1*</td>
<td>Patient ID</td>
<td>Numeric Value</td>
<td>&quot;It'd be in the same format for all of them [values in a variable].... not having disparate formats&quot; - Expert 2</td>
</tr>
<tr>
<td>2</td>
<td>Time of Observation</td>
<td>Datetime Type</td>
<td>&quot;It'd be in the same format for all of them [values in a variable].... not having disparate formats&quot; - Expert 2</td>
</tr>
<tr>
<td>3</td>
<td>Date of Birth</td>
<td>Date Type</td>
<td>&quot;It'd be in the same format for all of them [values in a variable].... not having disparate formats&quot; - Expert 2</td>
</tr>
<tr>
<td>4</td>
<td>Date of Birth</td>
<td>Normality</td>
<td>&quot;what do I need for a full observation to be good? ... all [values of] variables need to be normal.&quot; - Expert 1</td>
</tr>
<tr>
<td>5</td>
<td>Sex</td>
<td>Character Type</td>
<td>&quot;It'd be in the same format for all of them...not having disparate formats&quot; - Expert 2</td>
</tr>
<tr>
<td>6</td>
<td>Sex</td>
<td>Values fall in category (M=male) (F=female)</td>
<td>all values of Sex should fit available categories- Expert 1</td>
</tr>
<tr>
<td>7</td>
<td>Sex</td>
<td>Proportion of women is expected to be greater than men</td>
<td>&quot;women go to the doctors more frequently than men... if you look at a lot of clinical data, women outnumber men. - Expert 3, &quot;you'll have a few more women than men so maybe 60%, 65%.&quot; - Expert 4</td>
</tr>
<tr>
<td>8</td>
<td>Minutes Since Systolic</td>
<td>Numeric Type</td>
<td>&quot;It'd be in the same format for all of them [values in a variable].... not having disparate formats&quot; - Expert 2</td>
</tr>
<tr>
<td>9</td>
<td>Systolic Blood Pressure</td>
<td>Numeric Type</td>
<td>&quot;It'd be in the same format for all of them [values in a variable].... not having disparate formats&quot; - Expert 2</td>
</tr>
<tr>
<td>10</td>
<td>Systolic Blood Pressure</td>
<td>Systolic blood pressure falls in a range appropriate for risk calculation (90 mmHg - 225 mmHg)</td>
<td>Having a blood pressure below 90 is very unusual for adults, blood pressures below this are indicative of something life threatening and not appropriate for risk analysis. - Expert 3, &quot;in terms of the highest that I've seen...225.&quot; - Expert 4</td>
</tr>
<tr>
<td>11</td>
<td>Systolic Blood Pressure</td>
<td>Normality</td>
<td>&quot;what do I need for a full observation to be good? ... all [values of] variables need to be normal.&quot; - Expert 1</td>
</tr>
<tr>
<td>12</td>
<td>Minutes Since Diastolic</td>
<td>Numeric Type</td>
<td>&quot;It'd be in the same format for all of them [values in a variable].... not having disparate formats&quot; - Expert 2</td>
</tr>
<tr>
<td>13</td>
<td>Diastolic Blood Pressure</td>
<td>Numeric Type</td>
<td>&quot;It'd be in the same format for all of them [values in a variable].... not having disparate formats&quot; - Expert 2</td>
</tr>
<tr>
<td>14</td>
<td>Diastolic Blood Pressure</td>
<td>Diastolic blood pressure falls in a range appropriate for risk calculation (30 mmHg-150 mmHg)</td>
<td>&quot;I think 30 is a reasonable lower cut point&quot;- Expert 3, &quot;150 is probably enough&quot; - Expert 4, &quot;the lab does not always calculate LDL if triglycerides are above a certain number, because FHS does not require LDL a null value should not serve as an exclusion - Expert 3</td>
</tr>
<tr>
<td>15</td>
<td>Diastolic Blood Pressure</td>
<td>Normality</td>
<td>&quot;what do I need for a full observation to be good? ... all [values of] variables need to be normal.&quot; - Expert 1</td>
</tr>
<tr>
<td>16</td>
<td>Days Since Last Profile</td>
<td>Numeric Type</td>
<td>&quot;It'd be in the same format for all of them [values in a variable].... not having disparate formats&quot; - Expert 2</td>
</tr>
<tr>
<td>17*</td>
<td>LDL Cholesterol</td>
<td>Numeric Type</td>
<td>&quot;It'd be in the same format for all of them [values in a variable].... not having disparate formats&quot; - Expert 2</td>
</tr>
<tr>
<td>18*</td>
<td>LDL Cholesterol</td>
<td>Total cholesterol falls in a range appropriate for risk calculation (25-250 mg/dL) or null</td>
<td>&quot;I think 30 is a reasonable lower cut point&quot;- Expert 3, &quot;150 is probably enough&quot; - Expert 4, &quot;So for HDL... probably expected to be 10 or higher.&quot; - Expert 4, &quot;In terms of how high HDL can be you, I don't know, 100 or so, but just to be extreme you can say, maybe 200 wouldn't expect.&quot; - Expert 4</td>
</tr>
<tr>
<td>19</td>
<td>LDL Cholesterol</td>
<td>Normality</td>
<td>&quot;what do I need for a full observation to be good? ... all [values of] variables need to be normal.&quot; - Expert 1</td>
</tr>
<tr>
<td>20*</td>
<td>HDL Cholesterol</td>
<td>Numeric Type</td>
<td>&quot;It'd be in the same format for all of them [values in a variable].... not having disparate formats&quot; - Expert 2</td>
</tr>
<tr>
<td>21</td>
<td>HDL Cholesterol</td>
<td>HDL cholesterol falls in a range appropriate for risk calculation (10-200 mg/dL)</td>
<td>&quot;So for HDL... probably expected to be 10 or higher.&quot; - Expert 4, &quot;In terms of how high HDL can be you, I don't know, 100 or so, but just to be extreme you can say, maybe 200 wouldn't expect.&quot; - Expert 4</td>
</tr>
<tr>
<td>22</td>
<td>HDL Cholesterol</td>
<td>Normality</td>
<td>&quot;what do I need for a full observation to be good? ... all [values of] variables need to be normal.&quot; - Expert 1</td>
</tr>
<tr>
<td>23*</td>
<td>Total Cholesterol</td>
<td>Numeric Type</td>
<td>&quot;It'd be in the same format for all of them [values in a variable].... not having disparate formats&quot; - Expert 2</td>
</tr>
<tr>
<td>24*</td>
<td>Total Cholesterol</td>
<td>Total cholesterol falls in a range appropriate for risk calculation (75-500 mg/dL)</td>
<td>Consult literature to match highest and lowest that risk score was created from with a consensus. There are cases of cholesterol in the thousands but risk calculation is less appropriate for those with familial hyperlipidemia - Expert 4, Expert 3 &amp; Expert 2</td>
</tr>
<tr>
<td>25</td>
<td>Total Cholesterol</td>
<td>Normality</td>
<td>&quot;what do I need for a full observation to be good? ... all [values of] variables need to be normal.&quot; - Expert 1</td>
</tr>
<tr>
<td>26</td>
<td>Days Since Medication History</td>
<td>Numeric Type</td>
<td>&quot;It'd be in the same format for all of them [values in a variable].... not having disparate formats&quot; - Expert 2</td>
</tr>
<tr>
<td>27*</td>
<td>Anti-Hypertension Treatment</td>
<td>Binary Type</td>
<td>&quot;It'd be in the same format for all of them [values in a variable].... not having disparate formats&quot; - Expert 2</td>
</tr>
<tr>
<td>28*</td>
<td>Anti-Hypertension Treatment</td>
<td>Values fall in category (0 = no medication, 1 = medication)</td>
<td>&quot;the models take into account treatment with drugs.&quot; - Expert 3, Even if someone is on a hypertension medication for treatment of hypertension the medication is still reducing blood pressure and for the purpose of the equation, we still want to know any use of medication - Expert 3</td>
</tr>
<tr>
<td>29*</td>
<td>Days Since Current Smoker History</td>
<td>Numeric Type</td>
<td>&quot;It'd be in the same format for all of them [values in a variable].... not having disparate formats&quot; - Expert 2</td>
</tr>
<tr>
<td>30</td>
<td>Current Smoker</td>
<td>Binary Type</td>
<td>&quot;It'd be in the same format for all of them [values in a variable].... not having disparate formats&quot; - Expert 2</td>
</tr>
</tbody>
</table>
| 31 | Current Smoker | Values fall in category (0 = not current smoker; 1 = current smoker) | "I don't think you can categorize people as never I just don't because you're going to make the assumption that if they don't have smoking in their in their EHR that that then
<table>
<thead>
<tr>
<th>Rule Number</th>
<th>Description</th>
<th>Value</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>32</td>
<td>Patient ID, Time of Observation</td>
<td>Primary key; unique combination</td>
<td>the combination between patient ID and date time needs to be unique</td>
</tr>
<tr>
<td>33</td>
<td>Time of Observation, Date of Birth</td>
<td>Patient age is within the inclusion criteria for risk (39-79)</td>
<td>Ages outside the limit may be included in clinical practice but require a caveat. Ages outside the intent of the original score loses validation</td>
</tr>
<tr>
<td>34</td>
<td>Time of Observation, Date of Birth</td>
<td>Median age should be approximately 60 (within 10%)</td>
<td>Median and mean in the 60 range</td>
</tr>
<tr>
<td>35</td>
<td>Minutes Since Systolic, Minutes Since Diastolic</td>
<td>Diastolic blood pressure and diastolic blood pressure collected at the same time</td>
<td>if you're going to put these in a calculation, they should have been taken, at the same time</td>
</tr>
<tr>
<td>36</td>
<td>Minutes Since Diastolic; Time of Observation</td>
<td>Diastolic blood pressure is taken within a day of the observation</td>
<td>&quot;risk calculation should be taken at the time of a visit. Each visit should have a blood pressure associated with it. - Expert 3, &quot;in family medicine, we would expect that every single adult that comes through the doors has their blood pressure checked.&quot; - Expert 4, blood pressure has the most variance and should be considered an anchor measurement</td>
</tr>
<tr>
<td>37</td>
<td>Minutes Since Diastolic; Time of Observation</td>
<td>Diastolic blood pressure is taken within a day of the observation</td>
<td>&quot;risk calculation should be taken at the time of a visit. Each visit should have a blood pressure associated with it. - Expert 3, &quot;in family medicine, we would expect that every single adult that comes through the doors has their blood pressure checked.&quot; - Expert 4, blood pressure has the most variance and should be considered an anchor measurement</td>
</tr>
<tr>
<td>38</td>
<td>Systolic Blood Pressure, Diastolic Blood Pressure</td>
<td>Systolic blood pressure is always higher than diastolic blood pressure</td>
<td>&quot;The other issue is the systolic always has to be higher than the diastolic&quot;-Expert 3, &quot;you can never have a diastolic higher than a systolic.&quot; - Expert 4</td>
</tr>
<tr>
<td>39</td>
<td>Systolic Blood Pressure, Diastolic Blood Pressure</td>
<td>Systolic blood pressure is at least 10 points higher than diastolic blood pressure</td>
<td>the difference between systolic and diastolic should be &quot;greater than 10 and I wouldn't worry about how wide they are&quot;- Expert 4</td>
</tr>
<tr>
<td>40</td>
<td>Days Since Last Profile; Lipid Lowering Medication</td>
<td>If a person is on lipid lowering medication the profile should be within a year, else within 3 years</td>
<td>&quot;practice it's not unusual that two or three years go by without a you known lipid panel being we checked…. I think three years is the outside Window.&quot; - Expert 3, &quot;Also paying attention to whether or not they went on cholesterol lowering medication.&quot; - Expert 4</td>
</tr>
<tr>
<td>41</td>
<td>Hypertension Medication, HDL Cholesterol</td>
<td>If a person is on hypertension medication than they should not have higher than 50% drop between measurements; If not on hypertension medication they should not have higher than a 30% change</td>
<td>&quot;I wouldn't expect it to change more than like 30%. Its going to be very stable over time.&quot; - Expert 4, &quot;If someone goes on really high dose medicines, you might see a 50% drop.&quot;-Expert 4</td>
</tr>
<tr>
<td>42</td>
<td>LDL Cholesterol, HDL Cholesterol, and Total Cholesterol</td>
<td>LDL is calculated within 5% of the Friedewald Equation</td>
<td>Equation used to derive LDL from HDL, LDL and Triglycerides --&quot;it's called a Friedewald equation&quot; - Expert 3, &quot;Okay, Friedewald equation...it doesn't work very well for some people but it’s not bad.&quot; - Expert 4</td>
</tr>
<tr>
<td>43</td>
<td>Systolic Blood Pressure, Total Cholesterol</td>
<td>Blood pressure and cholesterol have a positive correlation</td>
<td>&quot;if they've got high blood pressure, chances are, they're probably going to have high cholesterol&quot; - Expert 2</td>
</tr>
<tr>
<td>44</td>
<td>Sex; Systolic Blood Pressure</td>
<td>Mean of blood pressure for men will be higher compared to blood pressure of women</td>
<td>&quot;generally, men are going to have higher blood pressure&quot; - Expert 2</td>
</tr>
<tr>
<td>45</td>
<td>Sex; Total Cholesterol</td>
<td>Mean of total cholesterol will be higher for men compared to women</td>
<td>men are going to have higher total cholesterol and LDL cholesterol - Expert 2</td>
</tr>
<tr>
<td>46</td>
<td>Sex; LDL Cholesterol</td>
<td>Mean of LDL cholesterol will be higher for men compared to women</td>
<td>men are going to have higher total cholesterol and LDL cholesterol - Expert 2</td>
</tr>
<tr>
<td>47</td>
<td>Sex; HDL Cholesterol</td>
<td>Mean of HDL cholesterol will be higher for men compared to men</td>
<td>&quot;women will probably have higher HDL cholesterol!&quot; - Expert 2</td>
</tr>
<tr>
<td>48</td>
<td>Days Since Current Smoker History; Current Smoker</td>
<td>If smoking status changes from a smoker to a non-smoker, 1 year should pass by between current and prior history</td>
<td>&quot;I seem to have seen two threshold six months and a a year&quot; - Expert 2, One year is a sufficient amount of time. - Expert 2 &amp; Expert 3</td>
</tr>
<tr>
<td>49</td>
<td>Patient ID</td>
<td>No null values</td>
<td>&quot;what do I need for a full observation to be good? …no null values&quot; - Expert 1</td>
</tr>
<tr>
<td>50</td>
<td>Date of Birth</td>
<td>No null values</td>
<td>&quot;what do I need for a full observation to be good? …no null values&quot; - Expert 1</td>
</tr>
<tr>
<td>51</td>
<td>Sex</td>
<td>No null values</td>
<td>&quot;what do I need for a full observation to be good? …no null values&quot; - Expert 1</td>
</tr>
<tr>
<td>52</td>
<td>Minutes Since Systolic</td>
<td>No null values</td>
<td>&quot;what do I need for a full observation to be good? …no null values&quot; - Expert 1</td>
</tr>
<tr>
<td>53</td>
<td>Systolic Blood Pressure</td>
<td>No null values</td>
<td>&quot;what do I need for a full observation to be good? …no null values&quot; - Expert 1</td>
</tr>
<tr>
<td>54</td>
<td>Minutes Since Diastolic</td>
<td>No null values</td>
<td>&quot;what do I need for a full observation to be good? …no null values&quot; - Expert 1</td>
</tr>
<tr>
<td>55</td>
<td>Diastolic Blood Pressure</td>
<td>No null values</td>
<td>&quot;what do I need for a full observation to be good? …no null values&quot; - Expert 1</td>
</tr>
<tr>
<td>56</td>
<td>Days Since Panel</td>
<td>No null values</td>
<td>&quot;what do I need for a full observation to be good? …no null values&quot; - Expert 1</td>
</tr>
<tr>
<td>57</td>
<td>HDL Cholesterol</td>
<td>No null values</td>
<td>&quot;what do I need for a full observation to be good? …no null values&quot; - Expert 1</td>
</tr>
<tr>
<td>58</td>
<td>Total Cholesterol</td>
<td>No null values</td>
<td>&quot;what do I need for a full observation to be good? …no null values&quot; - Expert 1</td>
</tr>
<tr>
<td>59</td>
<td>Days Since Medication History</td>
<td>No null values</td>
<td>&quot;what do I need for a full observation to be good? …no null values&quot; - Expert 1</td>
</tr>
<tr>
<td>60</td>
<td>Anti-Hypertension Treatment</td>
<td>No null values</td>
<td>&quot;what do I need for a full observation to be good? …no null values&quot; - Expert 1</td>
</tr>
<tr>
<td>61</td>
<td>Days Since Current Smoker History</td>
<td>No null values</td>
<td>&quot;what do I need for a full observation to be good? …no null values&quot; - Expert 1</td>
</tr>
</tbody>
</table>
Discussion

To pave the way toward more reliable CVD risk prediction through transparent and standardized EHR data quality assessments, we developed an explicit data quality assessment standard using the DataGauge process. This standard was composed of a data needs model that explicitly described the data and structure needed for CVD risk assessment using the Framingham 10-year hard coronary heart disease risk prediction model and a list of 61 data quality requirements to test EHR data’s fitness for CVD risk assessment. We also evaluated the standard’s coverage of data quality dimensions, knowledge domains needed for secondary use and levels of data granularity targeted by the DataGauge process, finding that all aspects were covered by our requirements. Assessing the coverage of dimension combinations from the 3x3 DQa framework, we found that all but one combination was covered by our requirements. Finally, we implemented our data quality standard to assess real-world EHR data for 456 patients, which identified 705 data quality issues in this limited test dataset. The core contribution of this work is to provide an initial, open data quality standard for CVD risk assessment using repurposed EHR data, previously unavailable, to our knowledge. This standard aims to serve as a starting point for the development of open, standardized and uniform data quality standards for the secondary use of EHR data. This is fundamental to the reliable secondary analysis of clinical data in the context of learning health systems.

The impact of EHR data on CVD risk assessment has been explored in the past, providing an initial understanding of the impact of EHR data’s imperfections on CVD risk scores. Specifically, the literature provides evidence that EHR risk assessment presents a great opportunity for the improvement of care, that there are still major challenges to ensuring that EHR-based CVD risk assessment is reliable, that EHR data quality and biases affect risk modeling, that the clinical context that impacts these data will impact the performance of these models and that EHR-specific bias seems to impact in risk assessment, but can be controlled for. Our paper extends this existing body of work in three ways. First, it provides an explicit data quality standard developed from expert knowledge for review, reuse and further development in future work. This opens up new avenues for research including collaborative standard development that will support both thorough “fitness for purpose-oriented” data quality assessments of EHR data, the transparent reporting of EHR data quality for secondary analyses and the support the reliable reuse of clinical data in the clinical setting. Second, our work also introduces the idea of integrating and cross-referencing health informatics data quality frameworks for the validation of data quality assessment designs. Though this has been explored as a means to harmonize data quality terminologies, using such integrations for data quality assessment work purposes has not been published in the past, to our knowledge. Finally, our findings show that it is possible to develop effective, purpose-specific data quality assessments based on expert knowledge following an explicitly-defined process, likely to improve the repeatability and reproducibility of data quality assessments, a standing problem in the field of EHR data quality assessment.

One of learning health systems’ aims is to identify and improve the care of patients at increased risk. Patients suffering from chronic disease such as those with diabetes (12-14% of U.S. adults, depending on criteria) have increased risk of CVD. The time-varying and risk-increasing nature of diabetes make this population particularly likely to benefit from EHR-derived, personalized CVD risk assessment over time. Still, personalized risk prediction models have not fully penetrated clinical practice. EHRs offer an opportunity to implement such models seamlessly in clinical practice leveraging large, longitudinal datasets for patients matched to specific populations. However, EHR databases present data quality issues that impact the reliability of these risk scores. Moreover, current data quality assessment methods are complex, challenging to implement, human-intensive and burdensome to implement. This reduces the likelihood and feasibility of conducting thorough data quality assessments for every secondary analysis of clinical data.

The routine reuse of clinical data for CVD risk assessment in the context of learning healthcare and clinical decision support may warrant the development of standardized data quality standards to increase the reliability of such analyses. For example, if EHR data were to be leveraged routinely to generate CVD risk assessment estimates at every visit, the repeated use would justify the overhead work of developing a data quality standard ahead of time. Approaches similar to the requirement testing used by the DataGauge approach appear to be scalable, providing further assets for implementation. Another way of increasing the efficiency of data quality standard development would be enabling quality standards (i.e., data needs models and corresponding requirements) sharing across research groups. This would save time and enable collaborative improvement of such standards. This is similar to the Automated Characterization of Health Information at Large-scale Longitudinal Evidence Systems (ACHILLES) module developed by researchers at the Observational Health Data Science and Informatics (OHDSI). However, this module only checks for general
(i.e., not purpose-specific) data quality and only focuses on the Observational Medical Outcomes Partnership (OMOP) data model. Further development along this approach for purpose-specific data quality standard development is likely to lead to fully-shareable quality standards for semi-automated quality assessments along with open, purpose-specific quality standards that could be developed, improved and refined via crowdsourcing the requirements development task within the informatics community. This is akin to the open-source development model. Despite the inherent potential of this approach, further development is needed to the fundamental infrastructure to enable and test its merits.

Our development and evaluation work presents four limitations mostly related to its preliminary nature. First, our data quality standard covers a single CVD risk assessment model (i.e., the Framingham 10-year hard coronary heart disease risk prediction model), despite a large number of alternative options. This risk model fits the needs of an initial pilot quality standard as it covers many factors included in other models but is also relatively simple compared to other models. Other models will be included in future work to expand the scope to CVD risk assessment at large. Second, we had access to a limited number of experts for the development of our data quality requirement standard. However, our development process covered all areas of expertise required for secondary analysis of clinical data according to the DataGauge guidelines. We also included an interdisciplinary expert to ensure that requirements affecting multiple areas of expertise would be covered. Still, we anticipate including a larger pool of single and multi-domain experts in future iterations of our CVD risk data quality assessment standard and hope to crowdsource the development of these requirements to the informatics community making routine secondary use of clinical data. Third, we evaluated our requirements’ validity according to two frameworks, despite many more being available in the literature. These frameworks were selected, on one hand, to align with our quality standard development method and test the internal validity using the data quality dimensions provided by the DataGauge methods and, on the other hand, to integrate our requirements with at least one existing published framework while testing their external validity. We selected the 3x3 DQa framework because its design dovetails with the DataGauge process, which frames the development of data quality requirements via a series of questions that integrate dimensions of data quality with levels of data granularity, whereas the 3x3 DQa framework integrates data quality dimensions with data constructs. Still, we will expand our integration of other data quality frameworks in future work. Finally, our initial data quality standard effectiveness evaluation was conducted using a limited set of patients from a single EHR system. The nature of this pilot work required a limited and focused dataset to test our standard’s capacity to identify data quality issues in real-world EHR data under restrictive conditions. Future work will address these limitations by expanding our focus to a larger set of patients representing the general population qualifying for CVD risk assessment and, subsequently, expanding to other EHR data sources.

Conclusion

We developed an explicit data quality assessment standard to support reliable CVD risk prediction via transparent and standardized EHR data quality assessments. The standard contained 61 data quality requirements that covered all data quality dimensions, knowledge domains needed for secondary use and levels of data granularity targeted by the DataGauge process and all but one of the 3x3 DQa framework quality dimension-data construct combinations. Our standard uncovered 705 data quality issues in a limited set of real EHR data for 456 patients. This explicit data quality standard for CVD risk assessment using repurposed EHR data can serve as a starting point for the development of open, standardized and uniform data quality standards for the secondary use of EHR data, which is fundamental to the reliable secondary analysis of clinical data in the context of learning health systems. Future work will include expansion of this standard to other risk prediction models as well as validation using other data quality frameworks and EHR databases.

Acknowledgements

This work was supported, in part by, funds provided by the University of North Carolina at Charlotte. This work was supported and used services and facilities, funded by the National Center for Advancing Translational Sciences (NCATS), National Institutes of Health (NIH) (UL1TR001420). This work was also supported by the NIH’s National Heart Lung and Blood Institute (NHLBI) grant R25 HL105400. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References


52. Curcin V. Embedding data provenance into the Learning Health System to facilitate reproducible research. Learning Health Systems. 2017;1(2).


Technology to Support Collaborative Dissemination of Research with Alaska Native Communities
Lisa G. Dirks, MS, MLIS & Wanda Pratt, PhD
University of Washington Information School, Seattle, WA, USA

Abstract
Marginalized communities often mistrust research due to a history of unethical practice and limited community engagement. Research community engagement is expected with Indigenous communities, but few empirical studies have explored engagement in results dissemination, let alone using technology. Studies on using technology to disseminate results focus on health and research professional audiences. This paper discusses Alaska Native stakeholder values on technology to facilitate collaborative results dissemination. In this formative study, six participants engaged in participatory design activities on collaborative results dissemination. Sketches and interviews were analyzed deductively using a value-based codebook. Study findings highlight the importance of community context and transparency. Contextual awareness includes understanding local culture and power dynamics, acknowledging the diversity of cultural practices within Alaska Native groups. Transparency is tied to clear communication: encouraging active dialogue and providing alternatives to communicate research. Technology that supports such collaborative dissemination could increase trust and improve adoption of research-recommended actions.

Introduction
Health research has impact when the people affected by health conditions can access and understand the research results, yet all too often, sharing coherent health research results with research participants and non-academic audiences is often neglected or done ineffectively. For many researchers, the primary method of disseminating research results is to submit manuscripts to peer-reviewed journals or conference presentations which limits their audience to academic scholars. Limiting dialogue between non-academic community participants and researchers, many of whom may be unacquainted with community perceptions of health or community systems of sharing information, decreases the likelihood that study results will be implemented at a community level. Collaboration between researchers and the communities that studies are conducted in can increase the likelihood that information is shared in more relevant community-centric ways. Studies suggest that research participants want to receive research results and propose that participants have a right to receive them for contributing to the research. Other than using traditional academic dissemination methods (i.e. journal articles, conference presentations), limited progress has been made to ensure health research results are returned to participants or otherwise shared with non-academic community audiences.

These dissemination limitations can lead to the spread of misinformation and mistrust, which is especially problematic for communities facing extensive social determinants of health. American Indian & Alaska Native (AIAN) people have encountered innumerable research injustices, such as being exposed to procedures without informed consent and having research results disseminated without community input or approval. For example, in the 1950’s Alaska Native people were unethically recruited into a study that involved ingesting radioactive iodine (I-131) at excessive doses without adequate consent and without appropriate documentation for follow-up which further motivates research mistrust. Oversights like these increase the need for transparency and accountability in research at all stages from project conceptualization to results dissemination. Moreover, AIAN people have been under-represented in national research initiatives—such as NIH’s All of Us research program—or have not been consulted about how best to implement research or disseminate actionable findings relevant to community context. Trust in health research influences trust in health services. Marginalized communities, such as AIAN people, are less likely to trust the scientific community as a result of unethical research practices. Egregious research violations have led to crucial research sovereignty in AIAN communities which has helped begin to mend their relationships with health research.

In the past several decades, research in AIAN communities has become much more collaborative or community driven. Despite this progress, only limited research has explored AIAN collaboration specific to results dissemination. Preliminary studies suggest that AIAN communities are interested in being engaged in the dissemination process. Previous research indicates that AIAN people want results that consider local context, appropriate language, information that is both practical to researchers and community members, and information that...

398
considers AIAN values, knowledge, and expertise. A collaborative approach to dissemination has potential to strengthen AIAN trust in research and enhance researchers’ awareness of community concerns and perceptions of research. Recently, AIAN-led organizations have seen success in informing AIAN communities about the risks and benefits of the COVID-19 vaccine. For example, the Urban Indian Health Institute conducted a US national survey with 1,435 AIAN respondents on community perspectives of the Covid-19 vaccine. Data from this survey have been used by the Urban Indian Health Institute and other AIAN serving organizations to develop social media campaigns and educational material about COVID-19 and vaccinations. A key finding from this survey reinforced the importance placed on community with 74% of respondents sharing a responsibility to their communities for getting vaccinated. This information has helped tailor COVID-19 prevention and education materials to community vs. individual benefit. Despite advances, limited empirical evidence is available about values AIAN stakeholders place on technology-facilitated collaborative dissemination let alone what they would want these technology tools to include. Further dialogue could encourage multi-directional learning and enhance partnerships leading to improved communication and trust in research promoting more participation in research where AIAN representation is now limited. New methods and technologies for respectful and culturally responsive research results dissemination are urgently needed both to help researchers engage diverse participants and to serve these communities better.

Internet-based technologies (i.e. social media, podcasts, websites, and mobile applications) have become more interactive and consumer-centered allowing for engagement and control over content that users interact with. These technologies increase the potential for conveying health-related research results to a broader audience. Moreover, these technologies have democratized access to research information, giving users the flexibility to manipulate how they see and share information including health research results. Bernhardt, Mays & Kreuter have coined the term “Dissemination 2.0” to represent the use of cooperative internet-based technologies to disseminate research products (e.g. results, educational content) claiming that it “represents an interactive approach of exchanging scientific evidence among collaborative members of a research-to-practice network that leverages their user-generated knowledge and harnesses their collective intelligence for increased effort and continuous improvement...” and takes advantage of the core foundations of Web 2.0 applications—collaboration, participation, multidirectional information exchange—to improve current dissemination activities and advance the translation of research to practice. This approach leverages relationships with healthcare practitioners, rather than average community members, but the core foundations of collaboration, participation, and multidirectional information exchange are potentially valuable for a wide range of communities. Preliminary studies on the use of internet-based and mobile technologies indicate a need to further explore these modes of research dissemination, particularly to community audiences.

Social media is recommended as a possible tool to broadly disseminate research results. Most studies investigating using internet-technologies to disseminate research results have focused on sharing results with academic and healthcare provider communities. Many have only reported web analytic data on website and social media reach, indicating increased access to their research content but did not provide much context about the user or their experiences. One study conducted with Australian, Indonesian, and Malaysian researchers asked about their social media research dissemination practices. Only 15% of the survey respondents reported using social media to share research results. In one web analytic study, Twitter was the most active platform for disseminating research articles; however, it is unclear whether these research articles were shared outside of academic networks or how tweeting them might impact science practice. Still, another survey conducted with clinical trial research participants suggested letters or fliers distributed via email and study website postings as most desirable, while Twitter, text messages, and conference calls were least desirable. It may be that using Twitter to share research results is more accessible to academic communities than it is for research participants.

Internet-based technologies have also been used to engage research-related dialogue between a research center and other researchers and community stakeholders. Soto et al. used Twitter to promote online community-engaged research training curriculum to both researchers and community members. Only 10% of the trainees were community members. While the researchers suggest social media-based community engagement is promising, there were important study limitations. Particularly, their inability to study demographic characteristics, including a user’s status as a community member vs. a researcher, means more research is needed to understand who the users of these strategies include. Feasibility studies on linking podcasts via social media (e.g. Twitter, Facebook) have also been explored as a novel form of communicating health research and facilitating engagement and dialogue. In one case, dialogue occurred through user comments and researcher replies on social media posts that linked research podcast recordings. As the study only examined web analytics and comments made on social media posts, the results of the study were limited to the comments made to the posts, and web analytics such as internet shares, and number
and duration of podcast plays. The authors suggest future studies are needed to explore how effective the podcasts were for reaching audiences, acquiring knowledge, and successfully communicating health topics content.  

This formative study explores health researcher and Alaska Native community stakeholder perspectives on technology-facilitated collaboration in the dissemination of health research results. Participants were engaged in a design process to develop low fidelity prototypes for community-researcher collaborative dissemination. This paper discusses participant preferences and values with respect to using technology to facilitate collaborative results dissemination.

Researcher Stance
Ms. Dirks positions herself as a mixed-heritage Alaska Native/White community-engaged researcher with community roots in an Alaska Native village in the Aleutian Islands. As an Alaska Native researcher, she has had experience being both a health and social sciences researcher as well as a community participant in health and social research studies. A significant amount of her research experience has been with Tribal communities in Alaska and elsewhere on topics relevant to community mental health and wellbeing. This dual role as a health services researcher and an Alaska Native person motivates her work and gives her unique perspective on both sides of the relationship between community and research.

Methods
This formative study was approved by the University of Washington Institutional Review Board. This study was conducted to develop a better understanding of: (1) stakeholder-defined risks and benefits in using technology to collaborate on health research results dissemination; and (2) stakeholder values related to collaborative research results dissemination. This exploration was framed using components of a value sensitive design (VSD) theoretical approach. VSD is an adaptable approach that engages moral and ethical considerations for technology design while also leaving room to explore other value considerations important to system stakeholders. It holds an “interactional position” which places attention on how people’s values determine how they interact with a technology even if it wasn’t designed for the explicit ways it is being used. It is an iterative process holistically integrating conceptual, empirical and technical aspects. This study involves the formative stages of a conceptual and empirical investigation of values relevant to community-researcher collaborative dissemination. Interest at this initial stage is to examine how well the conceptual value components compiled from the existing Indigenous community-engaged dissemination research literature align with the data contained in the participatory design activity artifacts created in this study. It is our intention to use data from this formative study to improve upon and iterate a value-based conceptual framework for collaborative dissemination considering an Alaska Native context.

Conceptual Investigation Methods
Reflecting on the flexibility of VSD, prior to empirical data collection, a literature review on researcher-community collaborative dissemination in Indigenous communities was conducted to create a list of value-based themes related to the design artifact data that were collected for this study (Table 1). These values were examined in relation to the design artifacts created during participatory design activities (See Figures 1 & 2 for example sketches).

Empirical Investigation Methods
A participatory design approach was used consisting of: (1) a list making activity in which participants were asked to write out a list of risks and benefits for using technology for Alaska Native community and researcher dissemination collaboration, (2) creation of a prototype sketch of the technology-based tool considering their ideas from the list-making activity; and, (3) scenario development which involved drawing or orating a story in which their prototype might be used. Using a deductive value-oriented coding manual, data obtained from sketches and accompanying scenarios created by participants were analyzed. The coding manual was created using conceptual definitions informed by existing literature. The literature that was used to inform definitions was isolated to community-engaged and collaborative dissemination-related health research specific to Indigenous communities. Table 1 provides a working list of these value-based conceptual definitions for collaborative dissemination.
Table 1. Preliminary conceptual value definitions and themes – Indigenous research results dissemination

<table>
<thead>
<tr>
<th>Value</th>
<th>Definition</th>
<th>Literature</th>
<th>Themes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Partnership</td>
<td>Arrangement where partners (community &amp; researchers) agree to cooperate to advance their mutual interests.</td>
<td>Blue Bird Jernigan et al., 2018</td>
<td>• Power dynamics&lt;br&gt;  o Internal (with community) &amp; external (researcher-community)&lt;br&gt;  o Individual vs. collective sense of community&lt;br&gt;  o Threat of technology on collectivism</td>
</tr>
<tr>
<td>Context</td>
<td>Considers the historical, physical, and social realities of the community in relation to the research study and research conducted in Native communities in general.</td>
<td>Legaspi &amp; Orr, 2007; McDonald et al., 2016; Timmons et al., 2007</td>
<td>• Epistemological nuance&lt;br&gt;  o Native ways of knowing differentiated from Western &amp; Alaska Native cultural groups&lt;br&gt;  o Geography&lt;br&gt;  o Rural geographic isolation&lt;br&gt;  o Pros and cons of technology use&lt;br&gt;  o Offer alternatives for in-person communication (pro)&lt;br&gt;  o Limited access to broadband (con)</td>
</tr>
<tr>
<td>Transparency</td>
<td>Conducting research in a way that it is easy for others to see what actions are performed. Transparency implies openness, communication, and accountability.</td>
<td>Bowen &amp; Martens, 2005; Elsabbagh et al., 2014</td>
<td>• Facilitates clear ongoing communication&lt;br&gt;  • Need technology skills to fully benefit&lt;br&gt;  • Limit technical jargon&lt;br&gt;  • Access guidelines &amp; restrictions</td>
</tr>
<tr>
<td>Dialogue</td>
<td>Considers the community as experts in their physical, social, and spiritual environment and meaningfully incorporates this knowledge in the results dissemination process.</td>
<td>Rivkin et al., 2013</td>
<td>• Mixed modes of communication&lt;br&gt;  o Text, audio, video, etc.&lt;br&gt;  • Synchronous/Asynchronous&lt;br&gt;  o Time for reflection (sometimes days in between needed)&lt;br&gt;  • Mutual educational benefit&lt;br&gt;  o Researchers learn from community &amp; community learn from researchers</td>
</tr>
</tbody>
</table>

Data collection process

A purposive sampling strategy consisting of email and word of mouth was used to recruit participants with strong connections to an urban Alaskan community. Participants were selected based on our prior knowledge of them having experience either being a health research participant or a health researcher. Inclusion criteria required participants to be 18 years or older and interested in exploring health research results dissemination processes in Alaska Native communities. Recruitment and data collection took place in May 2019 and early March 2020. Participants provided verbal consent. Design sessions were audio-recorded and transcribed, and extensive notes were taken throughout the process. All data collection took place in each individual participants’ private home. Participatory design sessions ranged from 30-60 minutes.

Participants

Six participants engaged with us during one of two time periods (May 2019 & March 2020). For the four who participated in May 2019, three self-identified as women, one as a man. Three self-identified as Alaska Native and one was White. One was a health researcher and the remaining three were Alaska Native community members living in an urban Alaskan community. For the two who participated in March 2020, one self-identified as a man and one as a woman. One identified as a non-Native health researcher with experience in Alaska and other American Indian communities. The other participant identified as an Alaska Native community member living in Seattle with roots in
a rural Alaskan community. Participant ages ranged from 37-68 years old. All participants had some experience in being research participants in previous health-related studies on topics in physical and behavioral health.

Findings
The lead author (Ms. Dirks) conducted a deductive analysis of session transcripts and written text on participatory design sketches and scenarios using a deductive value-oriented codebook consisting of the conceptual values noted in Table 1. Themes are highlighted below by these literature-informed conceptual values (partnerships, context, transparency, and dialogue). To distinguish between community and researcher participants, community participants are identified with the prefix “C” and researcher participants are prefixed with “D”.

Partnerships

Power Dynamics
Partnerships could benefit from awareness of power dynamics and engaging with various stakeholder groups in the community, not just those in a power position. This engagement should occur over time as a developing process, rather than as a product. Research results dissemination is often viewed as a product (i.e. manuscript, presentation). This limits its longevity. Seeing research as a process of ongoing communication, has impacts on partnership. Participants recommended youth as a community sector to involve in collaborative partnerships. Community participant C4 suggested that young people are often left out of decisions that have potential direct effects on them; they are often placed in a subordinate position to adults. Researcher participant R1 indicated that since youth will potentially be impacted by research in the long-term, they should be present to share perspectives about what problems they want to solve and what technologies they would want to interact with.

Individual vs. collective sense of community
Technology may pose possible threats to Alaska Native community norms of collectivism. A dissemination system has potential to cause isolation and individuation. Community participant C3 highlights this concern: “How did it work in the past for everyone to work together for communities to help sustain themselves? If it becomes more individualistic there is a breakdown. When does functional become dysfunctional? Would technology encourage the [individual] me-me-me attitude that was once [traditionally the Native collective] we-we-we?”.

Figure 1. Participant Example 1 Stakeholder Scenario Sketch.

Context

Epistemological nuance
The perception of Alaska Native communities’ epistemology is an important concept related to community values. Community participant C5 suggested that “[m]aybe sometimes Native people seem to be so indecisive because they are considering other parts of the community [in their decision making]. There is [a perception from outsiders that there is] no motivation to try to make things better”. This represents a tension between collectivist worldviews such as those found in many Indigenous communities in contrast to individualistic worldviews such as those most prevalent in mainstream United States communities. For this participant, misperceptions on decision-making processes can cause conflict if community context is misunderstood. Researcher participant R2 also stressed that each community has its own unique characteristics, especially within Alaska Native communities and stated that “sometimes we
[Alaska Native people] are grouped as one culture when we have so many cultural beliefs that vary even within our people”.

**Geography**

Geographic placement also has contextual importance. Rural Alaska Native communities are often difficult to get to. Transportation to and from communities becomes an important context value to consider. Some communities rely on air travel to get to and from their home location. This limits resources and opportunities for in-person collaboration due to constraints related to weather and airfare costs. Technology can be a benefit to communicating with people at a distance where costs of travel can range in the thousands of dollars for airfare alone. However, such an approach may also pose a problem for those communities that still have limited access to technology both peripheral and via infrastructure.

**Transparency**

Participants shared views on technology as both a facilitator and barrier to collaboration. C3 stressed importance for clear ongoing communication even before research results are ready to be presented. They viewed technology (i.e. websites) as a way for community members and researchers to maintain active communication. At the same time, C3 also viewed constraints of technology for transparency. For example, community members who have limited technology skills (e.g. Tribal Elders), may feel excluded from optimal collaboration making technology training or local facilitation an important consideration. Furthermore, transparency also includes using language and defining unfamiliar terms or norms, as stated by both researchers and community members. Researchers should limit technical research jargon and community members should define local terms and contextual factors that researchers may misinterpret. Community participants (C3, C5) also illustrated potential for mistrust in using technology to collaborate. They placed value on having guidelines on who has access, what it would be used for, and when people would have access.

![Figure 2. Participant Example 2 Stakeholder Scenario Sketch.](image)

**Dialogue**

Participants discussed benefits of technology for facilitating dialogue. To some participants, this dialogue does not need to be bi-directional oral communication but may benefit from having mixed modes of communication, such as the combined use of text, video, and audio. Researcher participant R4 discussed how dialog may benefit from being
both synchronous and asynchronous. In their scenario, community participants would meet as a local group independently to review non-technical research results that researchers would have shared with them before meeting in person. They stated that, in many Alaska Native communities, people need time to reflect before having discussion. A live videoconference could occur after the community group had time to reflect and deliberate on their own. This videoconference would be an opportunity for the community to ask researchers questions and share their response to the results and for researchers to orally articulate results and ask the community questions. The idea of mutual educational opportunities was also something to consider regarding dialogue. For example, researchers may learn from Alaska Native people about local knowledge, processes and interpretations, Alaska Native community members may learn more about science and research from the researchers. This organic communication is less likely to occur if a manuscript were the primary form of dissemination. Additionally, one community participant (C3) suggested that in-person communication would be ideal yet would prefer a technological interface for dissemination to reading a peer-reviewed article.

Implications for Technology Design
Empathy and sincerity towards the values of the communities we design with help us create technology systems that are likely to be more salient to their intended needs. Values can be difficult to isolate and often overlap and form networks showing interconnectedness. Co-designing with marginalized community participants can help expose values and how they are interrelated which can help equalize power dynamics between researcher and community member as well as expose community-level power dynamics that can unintentionally impact design. By being actively involved in technology design, community participants can highlight design flaws or features that do not align well with community collective norms. Active community participation can benefit design projects by having internal members of the community who understand cultural and epistemic nuances that may otherwise be missed by researchers with less familiarity with respective community context. As was discussed by one of our community participants, Alaska Native people have diverse cultural and linguistic backgrounds which make using a one-size-fits-all approach to participatory design for all communities a major challenge. According to the Bureau of Indian Affairs, there are presently 578 federally recognized tribes in the United States and many other tribal communities do not have federal recognition. Some of these communities have similar norms and practices but there is great variation across tribal communities in cultural protocols and traditions. Additionally, urban communities include representation from many of those tribes which makes it difficult to account for representation from each tribal community in those settings. Moreover, local participation and design facilitation, particularly in rural isolated communities, can benefit design by having support "on-the-ground" should travel or access to technologies become prevalent.

As with any participatory activity, it is important to understand the contextual nuances of conducting participatory design with a marginalized community, especially if the researcher is not already affiliated with the community. Although technology can support collaboration, it can also create barriers to collaboration that may not be exposed without understanding community norms. In our study, Alaska Native Elders or other community members with limited technology skills may have an opposite reaction to collaboration and instead feel isolated rather than more involved. An essential step to reveal these contextual distinctions is to spend time developing relationships with stakeholders and getting to know the community before initiating design activity. A proactive approach involving careful contextual analysis of the community, discovering potential problems, and developing and implementing strategies to address those challenges prior to starting design activities. Awareness of context also includes understanding local culture, including how to approach participants and show respect in culturally-specific ways which can be achieved by spending time with different stakeholder groups, many of whom may not be the design’s intended end user. Furthermore, developing relationships before design activities may also help uncover hidden agendas that may conflict with participatory activities.

To support partnership, context, transparency, and dialogue, co-design methods incorporated with Indigenous epistemological approaches will be added to our future research activities to promote more community power to the design process. For example, design processes could include a digital storytelling component that integrates oral traditions with technology design and research. Digital storytelling involves a process of creating short, personal stories that are told through a recorded, first-person voiceover, still and/or moving images, and music or sound. Digital storytelling connects and supports people in natural ways and is in alignment with AIAN cultural practices that promote values through sharing stories. Using narrative methods such as digital storytelling help deconstruct power dynamics between researchers and community members and are also respectful of rich oral histories and cultural practices in AIAN and other Indigenous communities. Integrating these approaches will also contribute to the community-engaged research approach by adding to the body of knowledge on engagement in community results
dissemination. Moreover, exploring research results dissemination from an Indigenous knowledge systems framework adds advantages to developing a culturally congruent conceptual framework that will increase sanguinity in health researchers’ communication with AIAN populations. This integration has the potential to generate innovative culturally responsive user-centered design methods that may be advantageous for participatory design activities with AIAN and similar communities with robust narrative values. The digital storytelling method will provide participants control over what they choose to share and how it is represented. Engaging participants in this way may increase parity by supporting their role as educators and owners of the knowledge they choose to share.

Conclusion

Conventional health research results dissemination approaches have often constrained marginalized communities’ research engagement. Creating interactive systems that enable community-researcher collaboration may benefit Alaska Native and other marginalized communities by encouraging active communication that is germane to their respective communities. Study findings highlight the importance of community norms and context in developing interactive dissemination systems for collaboration with Alaska Native communities. Awareness of context includes understanding local culture, including how to approach participants and show respect in culturally specific ways. We need to pay attention to power dynamics and dedicate time with different stakeholder groups, many of whom may not be the design’s intended end user. We also need to emphasize transparency with clear communication that promotes more active dialogue, potentially providing a variety of ways to communicate that would not otherwise be possible (i.e., mixed digital and analog media).

The findings from this study will be used to improve upon and iterate a value-based conceptual framework for technology-facilitated collaborative dissemination considering AIAN community contexts. The results of this and our future planned research will benefit improved communication and trust in health research in AIAN communities. Such increased trust will in turn improve the impact of health research and health services by ensuring that outcomes of research are more effectively disseminated to AIAN communities. It will support communities in understanding the research and contextual community-level relevance, potentially incorporating research recommendations and support for technology adoption. Such collaborative efforts could also encourage increased participation in national research programs by increasing marginalized groups’ trust in research. Similarly, this research builds connections between informatics researchers and AIAN communities urging further user-centered design collaborations. The conceptual framework, methodologies, and technology prototypes resulting from this and our planned future research could influence other historically marginalized communities by increasing their trust in and uptake of research results.

Acknowledgements

We would like to recognize the University of Washington GO-MAP fellowship program for the graduate funding support provided to Ms. Dirks during this project. We would also like to acknowledge the Alaska Native community members and health researchers who provided insight into the participatory design activities for this project. Professors Batya Friedman and David Hendry provided advice on VSD methodological approaches to use for this study and in future related research endeavors. Finally, we would like to acknowledge the iMed research group for their thoughtful discussions which prompted ideas on a previous research presentation of this formative study.

References


42. Del Gaudio C, de Oliveira A and Franzato C. The influence of local powers on participatory design processes in marginalized conflict areas. 2014, p. 131-139.

43. Hussain S, Sanders E and Steinert M. Participatory Design with Marginalized People in Developing Countries: Challenges and Opportunities Experienced in a Field Study in Cambodia. *International Journal of Design* 2012; 6: n/a.


Development of four electronic clinical quality measures (eCQMs) for use in the Merit-based Incentive Payment System (MIPS) following elective primary total hip and knee arthroplasty

Patricia C. Dykes, RN, PhD1,2, Mica Curtin-Bowen, BA1, Troy Li, BS1, Avery Pullman, BS1, Alexandra Businger, MPH1, Stuart Lipsitz, ScD1,2, Ania Syrowatka, PhD1, Michael Sainaire, MS1, Tien Thai, BS1, David W. Bates, MD, MSc1,2;

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

Abstract: The Centers for Medicare & Medicaid Services (CMS) supported Brigham and Women’s Hospital (BWH) Center for Patient Safety, Research, and Practice to retool one existing National Quality Forum (NQF) endorsed clinical quality measure (CQM) measure into an electronic clinical quality measure (eCQM) and develop three new eCQMs related to orthopedic care. This manuscript details the iterative process of measure development through environmental scans and stakeholder feedback prior to testing at two geographically different sites. The four measures under development are the: Risk Standardized Complication Rate (RSCR), Risk Standardized Venous Thromboembolism and Major Bleeding Rate (VTE/Bleeding), Risk Standardized Prolonged Opioid Prescribing Rate (POP), and the Risk Standardized Inpatient Respiratory Depression Rate (IRD).

Introduction:
Total hip arthroplasties (THA) and total knee arthroplasties (TKA) are common procedures that are associated with high costs and serious complications.1,2,3 Despite their frequency and cost, quality measurement tools to assess the safety of THA/TKA procedures are limited. Per the Health Services Advisory Group (HSAG) gap analysis reported in the Centers for Medicare & Medicaid Services (CMS) Quality Measure Development Plan, the orthopedics specialty has measurement gaps in areas throughout several CMS Quality Domains.4 Specifically, within the Effective Treatment/Clinical Care domain, there are no National Quality Forum (NQF) endorsed electronic clinical quality measures (eCQMs) related to orthopedic surgery outcomes.4 In addition to the lack of quality measurement tools, CMS is actively trying to reduce the regulatory burden that data collection and quality reporting in the Merit-based Incentive Payment System (MIPS) places on clinicians. MIPS provides performance-based pay to physicians based on the quality of care provided to their patients.5 To address the gaps in quality domains without adding to clinician documentation burden, CMS supported Brigham and Women’s Hospital (BWH) Center for Patient Safety, Research, and Practice to retool an existing clinical quality measure (CQM) and develop three new eCQMs for orthopedic care. eCQMs are tools used to measure surgical outcomes without adding an additional documentation burden for providers as they leverage routinely collected and readily available electronic health record (EHR) data to quantify rates at the individual clinician, clinician group, or hospital level. The measure development process focuses on the following four eCQMs at the clinician group level:
- Risk Standardized Complication Rate (RSCR) – retooled version of the existing NQF15506 and NQF34937 RSCR Clinical Quality Measures (CQMs)
- Risk Standardized Venous Thromboembolism and Major Bleeding Rate (VTE/Bleeding)
- Risk Standardized Prolonged Opioid Prescribing Rate (POP)
- Risk Standardized Inpatient Respiratory Depression Rate (IRD)
Over 139,000 TKA and 91,000 THA surgeries were performed in the United States in 2018 alone,1 and the volume of procedures being performed are projected to increase 71% for THA, and 85% for TKA by 2030.8 Considering the frequency of these procedures, these eCQMs are needed to address the current gaps in quality domains without imposing an additional documentation burden on clinicians.

Methods:
The development of the specifications for all four of the eCQMs was comprised of an environmental scan and stakeholder feedback which were used to validate the need for the measures, define inclusion/exclusion criteria, develop the risk adjustment models, and increase measure meaningfulness and feasibility.

Environmental Scan:
A structured literature review was conducted in collaboration with a Countway Librarian at Harvard Medical School. The reviews had two key objectives: 1) to identify and compile the current literature and practices published in the last five years, and 2) to determine patient-specific risk factors for each measure’s cohort. Additionally, a guideline review was conducted within Guideline Central. We searched the specialty area of “orthopedic surgery” for guidelines pertaining to THA or TKA that were published within the last five years. We then searched for general guidelines published in PubMed.

Existing measures:
A search for existing quality measures was conducted using the CMS Measure Inventory Tool (CMIT) and NQF Quality Positioning System (QPS) to locate all existing measures that are related to or competing with any of the proposed measures. A measure is defined by NQF as related if it involves either the same target population or the same concepts for measure focus, a measure is defined as competing if it involves both the same target population and the same measure focus.²

Technical Expert Panel:
Over the course of the measure development process, the BWH research team discussed the proposed measures with a Technical Expert Panel (TEP) comprised of surgeons, pharmacists, patient representatives, and measure development experts. The objective of the meetings was for the TEP to provide feedback on the specifications that would allow the proposed measures to meaningfully quantify adverse events following THA and TKA. Following the development process, the TEP also played an integral role in evaluating the field testing results for face validity and clinical meaningfulness. The BWH research team met with the TEP on five occasions throughout the development and testing process.

Structured Interviews:
The BWH eCQM development team partnered with an independent collaborator, Massachusetts Health Quality Partners (MHQP), to obtain stakeholder input on the proposed measures. MHQP used structured interviews with patients, providers, and payers to gather feedback and inform the decision-making process that led to the development of the eCQMs. To be interviewed, the following criteria had to be met:
- Patients: Adult, English-speaking patients who had undergone an elective primary THA and/or TKA at Brigham and Women’s Hospital or Brigham and Women’s Faulkner Hospital between January 1, 2017 and October 9, 2019. MHQP contacted these patients via email, and interested patients set up an interview.
- Providers: Any Mass General Brigham (MGB) orthopedic clinician was eligible to participate. This included surgeons, physician assistants, nurse practitioners, physical therapists, occupational therapists, and medical assistants. Recruitment was organized via email, and interested providers contacted MHQP to interview.
- Payers: All payers of Massachusetts health plans that provide coverage for orthopedic surgery were eligible to participate. MHQP contacted these payers via email, and interested participants set up an interview.

Public Comment:
The public comment period allows any interested parties to provide input on the measures via online surveys. Surveys were sent via email to stakeholders and relevant specialty organizations throughout the United States. The eCQM development team solicited comments via web posting for the RSCR measure and for the proposed risk adjustment model. A second round of public comments were solicited following measure development during the measure testing process.

Results:
Measure 1: Risk Standardized Complication Rate (RSCR)
Postoperative complications result in increased costs associated with THA and TKA. These complications affect the quality, and potentially quantity, of life for patients. Although complications following elective THA and TKA are rare, the impact can be devastating. Haro-Gómez et al. reports that the average cost for a hospital readmission after a THA is $17,103.² The costliest types of 90-day readmissions across the U.S. include infections (THA: $68 million, TKA: $89 million), acute cardiac events (THA: $33 million, TKA: $52 million), and acute vascular and thrombic events (THA: $13 million, TKA: $23 million), resulting in a large financial burden on the U.S. healthcare system annually.³ This proposed measure aims to quantify the complication rate at the clinician group level by harmonizing with and expanding upon the existing NQF1550 and NQF3493 measures.
Environmental Scan: 36 papers, two related quality measures\(^{6,7}\) (Table 1), and two practice guidelines\(^{10,11}\) (Table 2) met the environmental scan inclusion criteria and were used to inform inclusion/exclusion criteria, and measure specifications.

**Table 1: Guidelines used to inform RSCR measure development**

<table>
<thead>
<tr>
<th>Guideline</th>
<th>Developer</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinical Practice Guidelines on Surgical Management of Osteoarthritis of the Hip</td>
<td>American Academy of Orthopaedic Surgeons</td>
<td>2018</td>
</tr>
<tr>
<td>Clinical Practice Guideline on Surgical Management of Osteoarthritis of the Knee</td>
<td>AAOS</td>
<td>2016</td>
</tr>
</tbody>
</table>

Two related measures were identified in the environmental scan, which were the two measures which this proposed eCQM is aiming to retool (Table 2).

**Table 2: Existing measures related to the Risk Standardized Complication Rate eCQM**

<table>
<thead>
<tr>
<th>NQF #</th>
<th>Measure Steward</th>
<th>Measure Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>1550</td>
<td>Centers for Medicare &amp; Medicaid Services (CMS)</td>
<td>Hospital-level risk-standardized complication rate (RSCR) following elective primary total hip arthroplasty (THA) and/or total knee arthroplasty (TKA)</td>
</tr>
<tr>
<td>3493</td>
<td>CMS</td>
<td>Risk-standardized complication rate (RSCR) following elective primary total hip arthroplasty (THA) and/or total knee arthroplasty (TKA) for Merit-based Incentive Payment System (MIPS) Eligible Clinicians and Eligible Clinician Groups</td>
</tr>
</tbody>
</table>

**TEP:** The TEP discussed the RSCR measure on four occasions throughout the development process. In April 2019 when the measure was conceptualized using the same criteria as the NQF1550 and NQF3493, the TEP recommended that the RSCR measure be expanded from Medicare beneficiaries aged 65+ to all payers and all adults. This expansion was incorporated into all measures. In this meeting, a minimum of 25 procedures per clinician group was also defined for all measures. In September of 2019, the TEP recommended that BWH include both inpatient and outpatient surgeries and complications rather than only inpatient. Their rationale was that many serious TKA/THA complications are treated in emergency departments or in observation units.\(^{12}\) Restricting the measure to complications in inpatient settings would miss these events. In the May of 2020 and October of 2020 meetings, the development team shared testing results with the TEP. No changes were requested in these meetings.

**Structured Interviews:** Thematic analysis showed that patients (n=18) had mixed perceptions about the value of complication rate measures. Providers (n=10) shared that eCQMs, such as RSCR, add value by motivating improvements in practice and patient care and saving time. Providers were concerned that the eCQM could potentially incentivize providers to decline care for challenging cases. These concerns have been addressed through appropriate inclusion/exclusion criteria and a risk-adjustment.

**Public comment:** During the public comment period, the RSCR received 15 responses. Respondents shared that this measure would be ‘mostly useful’ to ‘very useful’ if implemented. 93.33% (14 out of 15) of respondents agreed that this measure could be used to facilitate quality improvements in THA/TKA outcomes.

**Final Measure Specifications:** The proposed RSCR measures complications at the clinician group level, includes all patients age 18 years and older from all payers, compared to Medicare Fee for Service (FFS) beneficiaries age 65 years and older, and includes procedures and complications documented in both inpatient and outpatient settings.

**Numerator:** The subset of patients from the denominator with any of the measure specified complications occurring during a period of 90 days following the THA or TKA procedure. The complications included in the measure must occur within the specified timeframe (Table 3) in order for the outcome to be attributable to the procedure and be included in the numerator. Complications and time frames were harmonized with NQF1550/3493\(^{13}\).

**Table 3: Complication timeframe following index admission**

<table>
<thead>
<tr>
<th>Complication</th>
<th>Timeframe (days)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acute myocardial infarction, pneumonia, sepsis</td>
<td>7</td>
</tr>
<tr>
<td>Pulmonary embolism, surgical site bleeding, death</td>
<td>30</td>
</tr>
<tr>
<td>Would infection/periprosthetic joint infection, mechanical complication</td>
<td>90</td>
</tr>
</tbody>
</table>

**Denominator:** All patients aged 18 years or older who received an elective primary THA or TKA procedure within the measurement year from all payers.

**Inclusion Criteria:**

---

410
• Aged 18 or older on the date of procedure
• Having a qualifying elective primary THA/TKA procedure (inpatient or outpatient setting)

**Denominator Exclusions:**
• Discharged against medical advice (AMA)
• Had more than two THA/TKA procedure codes during the index hospitalization
• Did not have a qualifying THA or TKA event.

These inclusion and exclusion criteria also apply to the other three measures under development.

**Measure 2: Risk-Standardized Major Bleeding and Venous Thromboembolism Rate (VTE/Bleeding) eCQM**

Major bleeding events following primary elective THA and TKA are associated with an increase in hospital length of stay, costs, and complications. \(^{14}\) Studies have estimated that about 4.7% of patients undergoing TJA would have symptomatic venous thromboembolism (VTE) without prophylaxis. \(^{15}\) There is general agreement within the orthopedic community that anticoagulant VTE prophylaxis is necessary after TJA, yet the use of anticoagulants puts patients at risk of experiencing a major bleeding event. \(^{16}\) Orthopedic surgeons face the challenge of balancing VTE prophylaxis against the associated risks of bleeding events, where insufficient anticoagulant use increases the risk for a VTE event, and a too high of a dose of anticoagulants increases the risk of a major bleeding event.

**Environmental Scan:** The BWH eCQM development team identified a total of 35 papers, five guidelines\(^{16-20}\) (Table 4), and six related measures\(^{21-26}\) (Table 5) to guide measure development. Review of existing guidelines and quality measures indicated that there are four related measures that examine major bleeding related to anticoagulants, and two that examine the incidence of VTE. No existing measures relating to VTE or major bleeding focus on THA/TKA as their target population, nor do these measures assess both VTE and bleeding.

**Table 4: Guidelines Used to Inform the VTE/Bleeding Measure**

<table>
<thead>
<tr>
<th>Guideline</th>
<th>Developer</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preventing Venous Thromboembolic Disease in Patients Undergoing Elective Hip and Knee Arthroplasty</td>
<td>American Academy of Orthopaedic Surgeons (AAOS)</td>
<td>2011</td>
</tr>
<tr>
<td>Reducing the Risk of Hospital-Acquired Deep Vein Thrombosis or Pulmonary Embolism</td>
<td>National Institute of Health and Care Excellence (NICE)</td>
<td>2018</td>
</tr>
<tr>
<td>Prevention and Management of Venous Thromboembolism</td>
<td>Scottish Intercollegiate Guideline Network (SIGN)</td>
<td>2010</td>
</tr>
</tbody>
</table>

**Table 5: Existing measures related to VTE or Bleeding events**

<table>
<thead>
<tr>
<th>NQF #</th>
<th>Measure Steward</th>
<th>Measure Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>450</td>
<td>Agency for Healthcare Research and Quality</td>
<td>Perioperative Pulmonary Embolism or Deep Vein Thrombosis Rate</td>
</tr>
<tr>
<td>376</td>
<td>The Joint Commission</td>
<td>VTE – 6: Incidence of Potentially Preventable VTE</td>
</tr>
<tr>
<td>2909</td>
<td>Agency for Healthcare Research and Quality</td>
<td>Perioperative Hemorrhage or Hematoma Rate</td>
</tr>
<tr>
<td>N/A</td>
<td>CMS</td>
<td>Adverse Drug Events for Patients Taking Anticoagulant Medications in an Ambulatory Setting</td>
</tr>
<tr>
<td>N/A</td>
<td>CMS/ Florida Medical Quality Assurance</td>
<td>Bleeding Outcomes Related to Oral Anticoagulants</td>
</tr>
<tr>
<td>N/A</td>
<td>CMS</td>
<td>Hospital-Harm – Medication related Bleeding (eCQM)</td>
</tr>
</tbody>
</table>

**Technical Expert Panel:** The TEP discussed this measure on three occasions during the measure development process. This measure was originally conceptualized to quantify the rate of post-surgical major bleeding events following THA and TKA, without incorporating VTE events. As pointed out by TEP members in the January 2020 meeting, measuring both events is key in ensuring a proper balance in the prophylactic regimens. The TEP unanimously agreed that expanding this measure from only major bleeding events to include both bleeding and VTE events would limit the unintended consequence of providers under-prescribing anticoagulants in order to avoid higher bleeding rates, which could in turn cause patients to develop VTE. The development team agreed that the expansion of the measure scope would increase measure meaningfulness and modified the eCQM to track both events. In the May of 2020 and October of 2020 TEP meetings, specification changes were made to improve face validity. Some changes made at the recommendation of the TEP included removing specific ICD codes from value sets, the addition of an anticoagulant use exclusion, and lowering the number of transfusions required to be included in the numerator from ≥2 to ≥1.
Structured Interviews: 23 patients, 15 providers, and 11 payers were interviewed regarding the measure concept and development. The overall consensus was that the proposed eCQM would be effective in providing clinicians with a representation of their anticoagulant prescribing practices through the assessment of both VTE and major bleeding rates following THA/TKA. The most common concern raised across interviews was that rates of VTE events may be very low, and that many surgical risk factors are associated with bleeding events aside from solely anticoagulant use. Additionally, even following the guidelines set forth by AAOS, there is no single standard of care for anticoagulation. Lastly, was also the concern that a patient’s lack of adherence to prescription regiments, resulting in an increased risk of bleeding or VTE event, would reflect poorly on the provider. The development team aimed to account for these problems, particularly those related to individual discrepancies, by applying appropriate inclusion/exclusion criteria to the eCQM, and through statistical risk-adjustment.

Final Composite Measure Specifications:
Numerator: Patients who develop a major bleeding and/or VTE event occurring from the date of the THA/TKA procedure to 35 days postdate of procedure.
Denominator: All patients, aged 18 years or older, who received an elective primary THA or TKA procedure and do not meet any exclusion criteria.

Measure Scoring: Clinician groups will be provided with a major bleeding rate and VTE rate based on the number of numerator events. The BWH eCQM team also developed a composite scoring system that takes into account the relative harm of each event. Based on the literature, bleeding events are more common but less harmful than VTE events. Using the Harm Weight Ratio from AHRQ PSI90, a VTE event is weighted as 2.7 times as harmful as that of a bleeding event (weighted at 1). However, while this scoring approach was approved by the TEP, current limitations in the Measure Authoring Tool (MAT) preclude this scoring method. The scoring approach used in the current measure configuration is limited to an equal weighting of Bleeding and VTE events. The development team will update the measure scoring system when MAT enhancements are made to support a weighted scoring system.

Denominator Exclusions (in addition to the exclusion criteria of the RSCR measure):
- With diagnosis codes for renal insufficiency within the 365 days prior to the THA/TKA procedure
- With diagnosis codes for chronic atrial fibrillation within the 365 days prior to the THA/TKA procedure
- With diagnosis codes for cancer within the 365 days prior to the THA/TKA procedure
- Who received prescription orders for anticoagulant medications 10-90 Days Prior to Surgery and meets the following criteria:
  - Patient who received an Anticoagulant Injection/Infusion
  - Patient who received a tablet (Oral) Anticoagulant order, Quantity > 1
- With VTE diagnosis code present on admission for index admission
- With major bleeding diagnosis code present on admission for index admission
- With diagnosis code for coagulation disorder within the 365 days prior to the THA/TKA procedure
- Who had additional surgery within 35 days from the elective primary THA/TKA

Measure 3: Risk-Standardized Prolonged Opioid Prescribing (POP) eCQM
Prolonged opioid use following TJA is a concern for patients due to the risk of developing an opioid dependence, overdose, or death. The goal of this measure is to assess the rate at which orthopedic clinician groups prescribe post-operative opioids to previously non-opioid exposed patients (patients who have not been prescribed opioids within 90 days prior to surgery) for an extended period (>42 days) following an elective primary THA/TKA.

Environmental Scan: 28 articles met the literature review inclusion criteria, as well as one post-operative opioid prescribing guideline and three relevant quality measures (Table 6 and Table 7).

Table 6: Guidelines used to inform POP measure development

<table>
<thead>
<tr>
<th>Guideline</th>
<th>Developer</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interagency Guideline on Prescribing Opioids for Pain</td>
<td>Washington State Agency Medical Directors’ Group (AMDG)</td>
<td>2018</td>
</tr>
</tbody>
</table>

Table 7: Existing measures related the POP eCQM
The TEP met to discuss this measure on four occasions. In April of 2019, the TEP discussed inclusion and exclusion criteria for the measure. The TEP and development team decided that documented diagnoses of Sickle Cell Disease, Cancer, and Opioid Use Disorder, as well as receiving hospice or palliative care, would be excluded from the measure, as these conditions and care types have different pain management and therapeutic goals in comparison to patients who do not have these conditions or are not in these care settings. The measure also focuses on patients not exposed to opioids and excludes patients with an opioid prescription within 90 days of the THA or TKA procedure. In January of 2020, the development team and the TEP agreed that the Washington State Guideline (2018) would serve as an effective basis for the proposed measure. In this meeting, the team began discussions about stratifying the hip and knee rates to reflect the differences in post-surgical opioid prescribing practices based on procedure type, where the knee arthroplasty population was showing much higher rates of extended use than the hip population in early beta testing. In the May of 2020 and October 2020 meetings, the TEP agreed to the stratification of hip and knee rates.

Structured Interviews: Qualitative interview results from MHQP showed that providers and payers noted benefits to the proposed POP measure including that the measure was: helpful in creating prescribing guidelines, increased likelihood of providers adjusting prescribing behavior, and that the correlation of opioid use duration and health plan costs encouraged behavior change among individual clinicians. Interview respondents saw the following potential risks to the measure: there were concerns about the ethics of restricting access to opioids and that the measure did not address variation in patient physiology, and issues with operationalizing opioid use. There was also the concern that hospital EHRs are not accurate sources of prescription information. These concerns are acknowledged as a limitation of the measure.

Final Measure Specifications:
Numerator: The subset of patients from the denominator who were prescribed post-operative opioids for >42 days after surgical discharge within the measurement year.
Denominator: The target population is all patients aged 18 years or older who received an elective primary THA or TKA procedure within the measurement year.
Denominator Exclusions (in addition to the exclusion criteria of the RSCR measure):
- The patient was prescribed opioids within the 90 days prior to the index admission
- The patient received a diagnosis of Opioid Use Disorder (OUD) within the 365 days prior to the index admission
- The patient had a Cancer diagnosis within the 365 days prior to the index admission or 90 days following discharge
- The patient had a diagnosis of Sickle Cell Disease within the 365 days prior to the index admission or 90 days following discharge
- The patient received hospice or palliative care within the 365 days prior to the index admission or 90 days following discharge
- The patient received an additional general or major surgery within 90 days following discharge
- The patient received a separate THA- or TKA-related procedure within the 90 days prior to the index admission or 90 days after hospital discharge

Measure 4: Risk-Standardized Inpatient Respiratory Depression (IRD) Rate eCQM
Background: The estimated incidence of inpatient respiratory depression (IRD) following elective primary THA and TKA ranges from 2.4%-25% in relevant literature due to a lack of defined metrics and universal definitions. Despite the lack of standards regarding diagnosis and coding, inpatient respiratory depression is a common postoperative pulmonary complication following TJA. Respiratory depression following elective primary THA and TKA puts patients at risk for brain hypoxia, anoxia, severe brain damage, cardiac arrest, and death. Respiratory depression following surgery is associated with longer hospital stays, and mortality rates 3.4 times higher than average. Analyses of malpractice claims estimate that 97% of inpatient respiratory depression cases were preventable with more frequent monitoring and early intervention.
Environmental Scan: 22 studies, three guidelines,39-41 (Table 8) and three quality measures42-44 (Table 9) were identified to inform measure development.

Table 8: Guidelines used to inform IRD measure development

<table>
<thead>
<tr>
<th>Guideline</th>
<th>Developer</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Management of Postoperative Pain: A Clinical Practice Guideline from the American Pain Society, the American Society of Regional Anesthesia and Pain Medicine, and the American Society of Anesthesiologists’ Committee on Regional Anesthesia, Executive Committee, and Administrative Council</td>
<td>American Pain Society, American Society of Regional Anesthesia and Pain Medicine, American Society of Anesthesiologists</td>
<td>2016</td>
</tr>
</tbody>
</table>

TEP: This measure was originally conceptualized as a Risk Standardized Opioid Induced Respiratory Depression Rate eCQM and was intended to measure the occurrence of respiratory depression attributable to post-surgical opioid use. The measure was respecified following advisement from the TEP, described below.

In the May of 2020 meeting, the BWH team proposed that an instance of opioid induced respiratory depression would be based on any of the three respiratory depression indicators: 1) diagnosis of respiratory depression-related outcomes, 2) mechanical ventilation or intubation procedure code, or 3) naloxone administration indicated for respiratory depression, in conjunction with administration of an opioid within 24 hours prior. The TEP shared with the team that opioid induced respiratory depression is a rare event following THA and TKA,34,35 and that it would be difficult to attribute respiratory depression to opioid use. The most reliable way to attribute opioid use to respiratory depression is the presence of naloxone administration and a documented diagnosis of RD; with this approach, rates would be very low, and the proposed measure would provide little opportunity for improvement. The TEP recommended that this measure be reframed to include all postoperative respiratory depression cases, including but not limited to cases induced by opioids. Following this advisement, the OIRD measure was renamed the Risk-Standardized Inpatient Respiratory Depression (IRD) Rate following THA/TKA. In the October of 2020 meeting, the TEP and the team discussed the results of alpha and beta testing of the IRD measure, the TEP was pleased with the overall progress of the measure.

Structured Interviews: In interviews with MHQP, patients, payers, and providers noted benefits of the proposed IRD eCQM including opportunities to compare clinician groups when deciding where to receive care or for health plans in the credentialing process. Interviewees also shared that this measure is not appropriate as a measure of care quality and could be imprecise (noted as a limitation of the measure). There were mixed perceptions about the level of attribution of a measure of inpatient respiratory depression, and that that primary care providers (PCPs), anesthesiologists, nurses, and other members of the orthopedic team may have some responsibility regarding respiratory depression.

Final Measure Specifications:
Numerator: the outcome for this eCQM is IRD occurring during the index hospital admission for TKA or THA. The outcome is dichotomous (i.e., yes or no). IRD is defined here as:
- Patient has a documented diagnosis of respiratory depression-related outcomes or respiratory failure
- Patient has a documented mechanical ventilation procedure code
- Patient has a documented intubation procedure code
- Patient has 3 or more SpO2 values ≤ 88 and ≥ 30 during their index admission following the procedure
  - Patient has 2 SpO2 values ≤ 88 and ≥ 30 that occur within 24 hours of each other during the index admission
  - The 88% threshold for oxygen saturation was chosen based off the literature in addition to incorporating the margin of error to arrive at a conservative value.45

Denominator: This eCQM includes adults 18 years of age or older, covered by any payer, undergoing inpatient elective primary THA or TKA.
Denominator Exclusions (in addition to the exclusion criteria of the RSCR measure):

- Patients who received an outpatient procedure

Risk Adjustment Model

Environmental Scan: Alongside the four measures under development, the environmental scan and stakeholder feedback was also used to inform the development of the risk adjustment model used in all measures. The risk-adjustment model was originally harmonized with the NQF1550 to include the following variables: number of procedures (1,2), procedure type (THA, TKA), and age. The environmental scan highlighted the following social determinants of health (SDOH) as relevant to include in the risk-adjustment model (beyond the variables harmonized with the NQF1550/3493): race, income (using ZIP Code information), smoking status, and language.

TEP: In the May of 2020 meeting, following the literature review and phone conferences with representatives from NQF in July of 2019 where NQF advised the expansion of the risk-adjustment model, the TEP confirmed the validity of adding the following SDOH to the risk adjustment model: race, household income (using ZIP Code information), language, and smoking status. The TEP also recommended the addition of body mass index (BMI) to the model.

Public Comment: Concerns raised during public comment were that the eCQMs may result in clinicians avoiding high-risk patients. These concerns were addressed in the inclusion of more SDOH variables in the risk adjustment model which will account for patients who are at higher risk for post-operative complications.

Final Risk Adjustment Model Specifications:
The following risk adjustment variables will be used for all measures: number of procedures (1,2), procedure type (THA, TKA), age, race, income (using ZIP Code information), smoking status, primary language, and BMI.

Exceptions:
- The POP measure does not risk adjust for procedure type (THA, TKA) as these rates are stratified
- Due to low expected rates of occurrences, the VTE/bleeding measure will not risk adjust for VTE events alone.
  The bleeding rate in addition to the composite bleeding and VTE score will still be risk-adjusted.

Expanding the risk-adjustment model provides a distinct benefit for the RSCR eCQM over the existing NQF1550 and NQF3493 measures and reduces the concern that providers would be penalized for taking on patients who are at a higher risk of post-surgical adverse outcomes across all measures.

Discussion:
Using a systematic, iterative approach based on an environmental scan, interdisciplinary stakeholder involvement, and specifications using standard codes and EHR data routinely available in different EHR systems, our team retooled one existing quality measure and developed three additional eCQMs for orthopedic surgery. The eCQMs measure key processes (opioid prescribing practices) and outcomes (risk standardized complication rates, VTE/bleeding rates, inpatient respiratory depression). The measures address a gap in orthopedic quality measurement, while leveraging data standards and routinely available EHR data to create measures designed to support benchmarking nationally while driving local quality improvement.

Existing literature on the development of eCQMs is limited. While there are published works on eCQM development in the fields of rheumatology, oral health, as well as published works on the testing of CQMs in the orthopedic field, there are currently no published papers on the development and specifications of eCQMs in orthopedics outside of those presented by the BWH development team. This manuscript builds on the body of literature regarding eCQM conceptualization and development and aims to inform future developers and stakeholders. The measure development team acknowledges several limitations:

Stakeholder Feedback: The BWH development team received feedback from healthcare providers, payers, and patients throughout the eCQM development process through public comment solicitation, structured interviews, and advisement from a technical expert panel. A limitation of this feedback was that it was conducted solely with English speaking adults, and did not include any non-English speaking patients, despite evidence that English proficiency is a predictor of surgical decision-making and outcomes. Future stakeholder engagement with non-English speaking patients is needed to assess measure validity across a larger subset of THA and TKA patients.

RSCR (and all eCQMs): The lack of data exchange between healthcare organizations limits the usefulness of eCQMs for provider groups that do not practice within an integrated care network. Complications for the RSCR eCQM extend 90 days post index admission. Therefore, complications that are treated outside the delivery network...
where the procedure was performed will not be available, making the RSCR appear lower than if all complications were included. **VTE/bleeding:** This measure was originally designed to use a weighted scoring method where VTE events would be weighted as 2.7, and major bleeding events would be weighted as 1.0, to reflect the severity of the adverse outcomes. This was not functional within the measure authoring tool (MAT), and thus, the outcomes are weighted the same despite their differences in frequency and lethality. **POP:** During the structured interviews, respondents saw potential risks to the measure including concerns about EHR systems not being accurate sources for prescription information. Other sources of opioid prescription information include state level databases which track a patient’s use, but the patient information for these databases is not presently accessible for use in eCQMs. **IRD:** The development team was initially interested in measuring the rate of inpatient respiratory depression that was attributed to postoperative opioid use. Due to the inability to reliably attribute respiratory depression to opioid use through EHR data and the low rates of OIRD, the measure was reformatted to measure inpatient respiratory depression without an opioid-induced component. In structured interviews, interviewees shared their concern that many factors contribute to respiratory depression and therefore the measure may be better used to evaluate health care systems, rather than orthopedic surgeons. They also reported that this measure may not be an appropriate measure of care since it could be imprecise. Due to the lack of defined metrics in diagnosing respiratory depression, the BWH development team believes that the use of EHR codes and the final measure specifications are the most feasible and reliable way to quantify inpatient respiratory depression rates at the clinician group level. Despite these limitations, our TEP members agreed that given the seriousness of respiratory depression when it does occur postoperatively, that this measure is meaningful and could lead to safer care for patients undergoing THA and TKA procedures. **Risk Adjustment Model:** During public comments, respondents shared that a potential unintended consequence for the measures would be that clinicians would not take on more complex patients who are more prone to post-surgical complications out of the concern over MIPS penalties. The development team addressed these concerns through the risk adjustment model, which aims to limit the individual characteristics of the patient so that adverse event rates are reflective of the clinician group’s performance, rather than the patient population. These measures have since been tested at two geographically different healthcare systems and risk adjusted following the measure development process. The next step in this research is to submit these proposed measures to the Measures Under Consideration (MUC) List and for NQF endorsement. These proposed eCQMs have the potential to highlight areas for quality improvement in orthopedics and provide clinician groups with metrics to quantify and compare their performance with other groups, all without imposing an additional documentation burden.

**Acknowledgements:** The BWH development team would like to thank Paul Bain, PhD, MLIS, of the Harvard Countway Library for his contributions to our environmental scan and literature review during measure development.

**References:**

Clinical Note Section Detection Using a Hidden Markov Model of Unified Medical Language System Semantic Types

Aaron S. Eisman, ScB1,2, Katherine A. Brown, RN, MSN1, Elizabeth S. Chen, PhD1,2,3, Indra Neil Sarkar, PhD, MLIS1,2,3,4

1Center for Biomedical Informatics, Brown University, Providence RI; 2The Warren Alpert Medical School, Brown University, Providence, RI; 3School of Public Health, Brown University, Providence, RI; 4Rhode Island Quality Institute, Providence, RI

Abstract

Clinical notes are a rich source of biomedical data for natural language processing (NLP). The identification of note sections represents a first step in creating portable NLP tools. Here, a system that used a heterogeneous hidden Markov model (HMM) was designed to identify seven note sections: (1) Medical History, (2) Medications, (3) Family and Social History, (4) Physical Exam, (5) Labs and Imaging, (6) Assessment and Plan, and (7) Review of Systems. Unified Medical Language System (UMLS) concepts were identified using MetaMap, and UMLS semantic type distributions for each section type were empirically determined. The UMLS semantic type distributions were used to train the HMM for identifying clinical note sections. The system was evaluated relative to a template boundary model using manually annotated notes from the Medical Information Mart for Intensive Care III. The results show promise for an approach to segment clinical notes into sections for subsequent NLP tasks.

Introduction

The electronic health record (EHR) is a rich source of biomedical data. With the passing of the Health Information Technology for Economic and Clinical Health Act of 2009 to incentivize the adoption of EHRs in medical practice, penetration of systems with core functionality, including clinician notes, has gone from about 27% to almost 100% of hospital systems over the past twelve years1. There is significant potential to leverage these data for the advancement of biomedical science and clinical medicine.

EHR data can be broadly separated into two categories. Structured data are stored as standardized values or codes. These include vital signs, laboratory results, medications, demographic information, problem lists, as well as diagnosis, procedure, and billing codes. The remainder of clinical data is stored in an unstructured form, such as in clinician narrative admission, discharge, and progress notes. Clinical notes contain a wealth of information that corroborates, contextualizes, and supplements what is available in a structured form. However, challenges surrounding the extraction of information from clinical notes (e.g., using natural language processing [NLP] techniques) have limited their use in large studies.

A notable portion of the EHR consists of unstructured free-text data. The clinical note is deliberately organized to enable efficient information retrieval by human readers. Each note category includes a series of expected section types. For example, the possible sections of a clinical note can be broken down into subjective information, objective information, and assessment and plan2. Subjective information is gathered using the interview and includes Medical History, Review of Systems, Medications, and Family and Social History. Objective information is obtained by the clinician’s in-person interaction with the patient or the ordering of diagnostic testing. In the clinical note, it comprises the Physical Exam and Labs and Imaging sections. The final section of Assessment and Plan represents clinicians’ summative conclusions about everything that precedes it and is where the next steps for the patient’s care are documented. Answers to individual clinical questions (e.g., the presence of chest pain), if present, can be reasonably expected to be confined to a small number of sections (e.g., Medical History or Review of Systems)3. While it is possible to further subdivide these section definitions4, in practice, those subdivisions are often combined inside a single paragraph or even within a single sentence (e.g., Assessment and Plan). Therefore, section identification at the level of granularity outlined is an effective first step to significantly increase the signal-to-noise ratio for downstream NLP tasks without significantly increasing section boundary ambiguity.

The method for note section identification for research purposes is almost exclusively performed using rule-based or machine learning approaches to identify institution-specific section headers along with formatting tokens like colons or bolded text that indicate the section header (e.g., “<b>Medications:<b>”)5. Prior work that modeled the sequence of note sections headings using a hidden Markov model (HMM) found that the majority of notes (67%) in some
clinical settings do not possess any section naming headers, and efforts to identify note sections were limited to the minority of notes that contained standardized section indicating tokens. Additional studies have reported high accuracy of granular section detection using section headers in notes to identify unlabeled sections. While this has been shown to be effective, it is challenged when either non-standard headers are used, the headers do not reflect the content that follows, or when seeking to develop generalizable approaches that can work across multiple clinical care settings with minimal training and configuration.

Generalizable approaches for identifying sections of the text using NLP tools include the identification of repeated words in adjacent text blocks and the identification of lexical chains using a thesaurus to link together related concepts. The brevity of clinical notes and lack of redundancy make direct application of these techniques that rely on repeated words unlikely to work. The most successful projects to automatically segment clinical text have used a shotgun approach of attempting to identify section breaks based on whitespace and variations in section headers along with markers for header definition such as colons and bolded type with good but not perfect accuracy. The success of this is highly dependent on the institution and EHR system-specific structure of clinical notes.

Inspired by these concepts, this study set out to use the Unified Medical Language System (UMLS) Metathesaurus to identify semantic concepts within sections of clinical text to characterize the distribution of substantive content of note sections. UMLS concepts were chosen as a domain-appropriate knowledge structure to serve as a proxy for informative information contained within note sections. The concept-based semantic types further reduced data dimensionality such that they could be modeled as a Markov process. It was thus hypothesized that a heterogeneous HMM using semantic type distribution and section transition probabilities would provide a generalizable note section detection algorithm.

Methods

Training and testing sets of admissions notes were randomly sampled from a publicly-accessible EHR dataset. The notes were then manually labeled for component sections based on content. These annotations were used to calculate the probability of transitioning from one section to another. The note text was also processed by the MetaMap NLP system to identify UMLS semantic types. Probabilities of observing the semantic types were calculated for each note section type. A subset of semantic types were selected for inclusion based on their ability to uniquely identify a note section type. Collectively, these defined a heterogeneous HMM that was trained and cross-validated on the first set of notes and tested on the second. The performance of the HMM was evaluated and compared to a template to identify note sections. The overview of these steps are outlined in Figure 1.

MIMIC-III Database

Clinical notes from the Medical Information Mart for Intensive Care III (MIMIC-III) database were used for this analysis. MIMIC-III is a publicly available de-identified dataset of intensive care patients treated at Beth Israel Deaconess Medical Center from 2001 to 2012 and includes approximately 60,000 critical care admissions. A set of 100 notes that used the single most common admission note template were randomly selected as a training set. The single most common template was selected for training in order to allow out-of-sample evaluation of different note templates to demonstrate how sensitive the HMM method is to template changes. A testing set of 100 notes was

![Image of Figure 1. Overview of the Study](image-url)
selected that comprised a variety of templates. Many of these contained no relevant changes, and those that had changes often had differences in one or two headers (e.g., “HPI:” vs “History of Present Illness:”).

Manual Annotation

A reference standard was developed to train and test the automated identification of note sections. Using written guidelines, two human annotators (KAB and INS) with clinical and informatics expertise independently identified note sections in the same two hundred clinical notes with standardized annotations at the beginning of new sections. Disagreement between the two annotators was adjudicated by a third annotator (ASE). Up to seven different section types were identified in each note. (1) Medical History was defined as including the chief complaint, history of present illness, past medical history, and review of systems if continuous with the rest of medical history. (2) Medications also included allergies if the two appeared contiguously in the note. (3) Review of systems was labeled separately if not adjacent to a medical history section. (4) Family and social history were combined into a single section type due to the fact that they are sometimes combined into a single line of text making it impossible to distinguish them at a line-level granularity. (5) Physical exam was defined to include vital sign and other “flowsheet” information in addition to the standard physical exam. (6) Labs and Imaging including any objective testing information results included in the note. (7) Assessment and Plan was defined as a summative analysis of the patient’s status at the time of note writing followed by a description of next steps for medical care.

MetaMap Text Processing

MetaMap is an NLP tool maintained by the National Library of Medicine for the processing of text for the identification of biology and medicine UMLS concepts. The full set of notes was processed using the MetaMap Docker image in batch. MetaMap was configured to use the NLMSubSyn vocabulary. UMLS concept unique identifiers (CUI) and their associated semantic types were tracked, along with their associated line number from the original note text file. Absolute and relative frequencies of semantic types identified within each section type were calculated for the notes in the training set. A subset of semantic types (n=24) were selected for the HMM that (1) occurred at least 1.5X more frequently in a single section type compared to the note as a whole, and (2) occurred in at least 90% of the sections of the type for which it is overrepresented.

Hidden Markov Model

A hidden Markov model (HMM) is a simple dynamic Bayesian network of a Markov process with underlying unobserved states that influence the observed output. There are three main parameters of an HMM. The first is an initial probability array, \( \pi \), which contains the probability of starting in each of the possible unobserved states. The second is a state transition probability matrix, \( A \), which contains the probability of changing from each possible unobserved state to another unobserved state (including remaining in the same state) after each observation. Finally, there is an observation probability matrix, \( b \), that contains the probabilities of each observation for each state. In the context of the task of note section identification, the note sections are the unobserved hidden states and the UMLS semantic types associated with MetaMap identified UMLS CUIs are the observations. The model was constrained to only allow for the identification of a state transition after the end of a new line in the note. This was accomplished by having an identity matrix, \( a_2 \), that represents the probability of state changes between observations on a single line of note text.

The parameters for the HMM were trained using the manually annotated clinical notes from the training set. The initial state probability array, \( \pi \), was defined as the proportion of notes that started with each section type. State transition probabilities for \( a \) were empirically determined on a per line basis. For each note, the transition probabilities were determined for all sections within that note with each new line being considered an opportunity for state transition. The arithmetic mean of each state transition probability was then calculated across all notes and normalized to sum to a total probability of one. Observation probabilities were also calculated within each note for each section. They were defined as the proportion of each semantic type divided by the total semantic types in each section. The observation probabilities matrix was then calculated as the mean of the probabilities calculated for each note.

Template Boundary Model

A template boundary model (TBM) was developed using the same most common admission note template applied in the training set. This model looked for the words or phrases that indicate the start of each of the corresponding note sections using regular expression matching of standardized section headers (Table 1). In the event that consecutive template sections were assigned to the same “mapped section” those were considered a single merged section.
Table 1. Template Boundary Model

<table>
<thead>
<tr>
<th>Template String</th>
<th>Mapped Section</th>
<th>Template String</th>
<th>Mapped Section</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chief Complaint:</td>
<td>Medical History</td>
<td>Family history:</td>
<td>Family and Social History</td>
</tr>
<tr>
<td>HPI:</td>
<td>Medical History</td>
<td>Social history:</td>
<td>Family and Social History</td>
</tr>
<tr>
<td>Allergies:</td>
<td>Medications</td>
<td>Review of systems:</td>
<td>Review of Systems</td>
</tr>
<tr>
<td>Infusions:</td>
<td>Medications</td>
<td>Flowsheet Data:</td>
<td>Physical Exam</td>
</tr>
<tr>
<td>Other ICU medications:</td>
<td>Medications</td>
<td>Vital Signs:</td>
<td>Physical Exam</td>
</tr>
<tr>
<td>Other medications:</td>
<td>Medications</td>
<td>Labs / Radiology:</td>
<td>Labs and Imaging</td>
</tr>
<tr>
<td>Past medical history:</td>
<td>Medical History</td>
<td>Assessment and Plan:</td>
<td>Assessment and Plan</td>
</tr>
</tbody>
</table>

TBM was also applied to the testing set. The results of this model were used to contextualize the evaluation of the HMM method and allow a comparison of each method’s sensitivity to small changes in the template.

Evaluation

Five-fold cross-validation of the HMM method was performed by randomly separating the data into five equal bins. The model was trained on four of the bins (i.e. HMM parameters were empirically determined) and subsequently tested on the remaining bin. Using this process, each original note was labeled by the HMM exactly one time. This process was performed once on the training set. The content of a note section in the reference standard was defined as the line number occurrence of an annotated section up to but not including the line of the next label or the end of the document. The average of the parameters from the five-fold cross-validation was carried forward as a single model applied to the testing set.

The performance of the identified note section boundaries was measured by the sensitivity and specificity of the content those boundaries define. The model was evaluated as the number of UMLS concepts (e.g., a proxy for the interesting data) captured by the identified section boundaries. Concepts for a given section were considered true positives if they were within the manually annotated and model boundaries for the section. False positives were defined as concepts inside the model boundaries but outside the annotated ones. Concepts outside of the model boundaries but within the annotated boundaries were considered false negatives. UMLS concepts outside of both the model and annotated boundaries were considered true negatives. Full confusion matrices were calculated using these definitions for the HMM and the TBM for both the training and testing datasets and then used to calculate sensitivity, specificity, precision, recall, and F1-score.

Results

Note Annotation Observations

Two hundred clinical notes were manually annotated into 1,336 distinct sections based on the annotation guidelines and reference standard as described above. Overall, the notes were organized and formatted in a chronological fashion from the time prior to the note being written to planned future clinical care. The notes were formatted in the original dataset such that lines were limited in length to about twelve words. As a result, paragraphs of text and even individual sentences were often interrupted by a newline character. In addition, many parts of the notes did not contain complete sentences. For example, lists were commonly employed to record medications, symptoms, and medical history not directly pertaining to the chief complaint. Each note typically began with a history of present illness and medical history. This was followed by varying combinations of Medications, Review of Systems, Family and Social History, Physical Examination, and Labs and Imaging results. Notes almost always ended with an assessment of the patient’s status and a plan going forward, including medications, testing, and follow-up care. Most of these sections contained discrete labels within the text.

There was variation in how sections were labeled, and at times they were entirely mislabeled. For example, the most commonly mislabeled sections included Medical History, and Family and Social History. These section headers were often listed out, and then the content for all three sections was merged together in the lines that followed the header labels. Another less frequent example included content from distinct sections that were embedded within one another.
(e.g., home medication list followed by medical history/surgical history, and ending with a list of more or duplicate home medications). Other observations throughout manual annotation included extra text that did not pertain to a particular section or additional notes as an addendum added at the end that included brief summaries of Medical History, Physical Exam, and Assessment and Plan written by another clinician (e.g., specialist, attending, or medical student). The annotators labeled sections based on content and not the section headers. If a section header corresponded to the following content, it was marked as the start of the section. As expected, it was evident that the unique characteristics, language, content, length, and formatting of the clinical note varied based on individual provider and specialty.

Transition State Probabilities

A transition state probability matrix was calculated using training data for each fold in the cross-validation. This was repeated five times. Mean transition probabilities after a newline character are diagrammed as a Markov process (Figure 2a). For all sections, it was found to be most common to stay within that section. Beyond that, Medical History can transition into Medications or Family and Social History. Family and Social History can be followed by Medications or Physical Exam. The remaining sections of Review of Systems, Physical Exam, Labs and Imaging, and Assessment and Plan always follow in sequential order. An example portion of the heterogeneous HMM, including both observations and hidden states, is shown in Figure 2b. Observations are UMLS CUIs converted to UMLS semantic types (e.g., Metronidazole → Antibiotic). State transitions are only allowed between observations separated by a newline character with probabilities shown in Figure 2a. Between observations without a newline character, the hidden state is maintained. The most likely path is determined using expectation maximization. A hypothetical example of this is bolded in Figure 2b as Metformin → Pharmacologic Substance (Medications) to Tobacco → Plant (Family and Social History). The most common progression of note sections is visualized as an alluvial diagram (Figure 2c). One-half of the notes followed the most common section progression of Medical History, Medications, additional Medical History, Family and Social History, Physical Exam, Labs and Imaging, and then Assessment and Plan. About one-quarter of the notes also had a Review of Systems section, most commonly between Family and Social History and Physical Exam.

Semantic Type Feature Distribution

The 24 UMLS semantic type features were calculated for all 1,366 note-sections in the reference standard. Semantic type distributions by note section type are depicted in Figure 3. The analysis of semantic type distribution represented by different sections was in line with expectations. Medications, Family and Social History, Physical Exam, Labs and Imaging, and Review of Systems all contain distinct information represented by specialized language that was expected to generate unique probability distributions.

It was found that more than two-thirds of the identified concepts within the Medications sections were “Pharmacologic Substance”, “Organic Chemical”, or “Antibiotic”. Family and Social History was dominated by “Organic Chemical”, Pharmacologic Substance”, and “Plant”, all referring to substance use habits (e.g., Tobacco). In addition, “Family Group” is significantly overrepresented, likely reflecting the common format for family history (e.g., “Father had a myocardial infarction at 64 years old”) and description of living arrangements, marital, and family status that are commonly included in Social History. Review of Systems is expectedly overweight in “Body System” and “Sign or Symptom”. Similarly, an outsized proportion of “Body Location or Region” within the Physical Exam sections is indicative of the kinds of notations that are made about the physical exam, which represents in-person observations about a patient’s body. Finally, the ambiguity presented by Medical History and Assessment and Plan was an expected challenge, but the latter includes an expectedly higher proportion of “Idea or Concept” words (e.g., “probably” and “consistent with”) that reflect the author’s thought process when assessing the patient’s condition. In addition, differentiating these two sections is accounted for by the transition probabilities matrix of the HMM.
Figure 2. Note Section Transitions. a) Markov Process, b) Heterogeneous HMM, c) Alluvial Diagram
Hidden Markov Model Evaluation

Both the HMM and TBM were used to identify note sections in two-hundred admission notes. Training was performed on one-hundred notes that used the single most common template within the MIMIC-III database. The performance of the HMM on these notes compared to the TBM is reported as five-fold cross-validation of the HMM in Table 2. Overall, both models performed well. The HMM captured greater than 85% of the semantic types in all sections, including better than 92% in Review of Systems, Physical Exam, and Assessment and Plan with the exception of Labs and Imaging, which performed relatively poorly (70%). The specificity of the HMM was at least 93% across all note section types. The HMM significantly outperformed the TBM in identifying Medical History concepts (88% vs. 75%) and significantly underperformed the TBM in identifying Labs and Imaging concepts (70% vs. 99%). The HMM modestly underperformed the TBM (Δ ≤ 10%) in the remaining note section types. Specificity was excellent for both and only significantly differed for two section types where the HMM outperformed the TBM for Family and Social History (99% vs. 93%) and underperformed the TBM for Assessment and Plan (93% vs. 100%).

The performance of the HMM compared to the TBM on the testing set is reported in Table 3. The HMM achieved a sensitivity of at least 77% across all note section categories. It performed at 84% for Medical History and Review of Systems and at least 93% for Physical Exam and Assessment and Plan. The sensitivity of the HMM outperformed the TBM in every note section category except Family and Social History; however, this was achieved by the TBM with only 20% precision. The HMM achieved very high specificity (≥ 94%) for all note section types, comparable to the TBM.

Between the training and testing sets, the HMM had an average sensitivity performance reduction of less than 3% per section type compared to a 14% degradation in the TBM. For the HMM, no single section degraded more than 8% (Family and Social History and Review of Systems), and the worst-performing section for the HMM in the testing set, Labs and Imaging, improved by 7%. Overall, the HMM performance was consistent between training and testing sets when compared to the TBM, where every section had reduced sensitivity of at least 10%, with the exception of Medical History, which was nearly unchanged.
Table 2. Five-Fold Cross-Validation (100 Single Template Notes)

<table>
<thead>
<tr>
<th></th>
<th>Specificity</th>
<th>Precision</th>
<th>Recall/Sensitivity</th>
<th>F₁</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HMM</td>
<td>TBM</td>
<td>Δ</td>
<td>HMM</td>
</tr>
<tr>
<td>H</td>
<td>0.99</td>
<td>0.97</td>
<td>0.01</td>
<td>0.94</td>
</tr>
<tr>
<td>M</td>
<td>0.99</td>
<td>1.00</td>
<td>-0.01</td>
<td>0.61</td>
</tr>
<tr>
<td>F</td>
<td>0.99</td>
<td>0.93</td>
<td>0.06</td>
<td>0.81</td>
</tr>
<tr>
<td>R</td>
<td>0.99</td>
<td>0.99</td>
<td>0.00</td>
<td>0.67</td>
</tr>
<tr>
<td>P</td>
<td>1.00</td>
<td>1.00</td>
<td>0.00</td>
<td>0.98</td>
</tr>
<tr>
<td>L</td>
<td>0.99</td>
<td>1.00</td>
<td>-0.01</td>
<td>0.86</td>
</tr>
<tr>
<td>A</td>
<td>0.93</td>
<td>0.99</td>
<td>-0.06</td>
<td>0.87</td>
</tr>
</tbody>
</table>

HMM - Hidden Markov Model; TBM - Template Boundary Model; Δ - Difference (HMM-TBM)

Table 3. Out of Sample Testing (100 Notes with Minor Template Modifications)

<table>
<thead>
<tr>
<th></th>
<th>Specificity</th>
<th>Precision</th>
<th>Recall/Sensitivity</th>
<th>F₁</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HMM</td>
<td>TBM</td>
<td>Δ</td>
<td>HMM</td>
</tr>
<tr>
<td>H</td>
<td>0.96</td>
<td>0.94</td>
<td>0.02</td>
<td>0.86</td>
</tr>
<tr>
<td>M</td>
<td>0.96</td>
<td>0.99</td>
<td>-0.03</td>
<td>0.61</td>
</tr>
<tr>
<td>F</td>
<td>0.99</td>
<td>0.91</td>
<td>0.08</td>
<td>0.69</td>
</tr>
<tr>
<td>R</td>
<td>0.99</td>
<td>0.99</td>
<td>0.00</td>
<td>0.74</td>
</tr>
<tr>
<td>P</td>
<td>0.99</td>
<td>1.00</td>
<td>-0.01</td>
<td>0.95</td>
</tr>
<tr>
<td>L</td>
<td>0.98</td>
<td>1.00</td>
<td>-0.02</td>
<td>0.83</td>
</tr>
<tr>
<td>A</td>
<td>0.94</td>
<td>0.99</td>
<td>-0.05</td>
<td>0.91</td>
</tr>
</tbody>
</table>

HMM - Hidden Markov Model; TBM - Template Boundary Model; Δ - Difference (HMM-TBM)

Discussion

Admission notes from MIMIC-III represent real-world examples of clinical notes for ICU patients. The manual annotation process clearly demonstrated that while these notes possessed the expected structure that domain experience suggested, they were far from ideal. The clinical note text was commonly untidy and poorly arranged. For example, many of the Labs and Imaging sections were intended to be in a tabular format but were generally misaligned. In addition, some sections were routinely mislabeled or labeled without any relevant content to follow. In addition, there were frequent line breaks that interrupted sentences and paragraphs. These are likely artifacts of data storage processing and the result of irregular use of note templates. All of these findings reestablished the importance of note section detection informed by content rather than formatting tokens.

It is important to acknowledge that the results presented here represent a single note type tested from a single institution. Furthermore, all patients within MIMIC III were treated in an ICU, which likely both increases the complexity of the recorded information in the associated clinical notes but also necessitates structure. Significantly shorter notes that contain no underlying structure are unlikely to benefit from section detection using the methods described in this study.

The HMM method performed favorably to the TBM in several important ways. The goal of this comparison was to evaluate the model robustness to support portability to changes in templates over time or to other electronic health record systems. The most dramatic example was with Family and Social History. In this dataset, while the headings...
were consistently used, they frequently did not appropriately precede relevant content. Therefore, while the TBM appropriately found the Family and Social History section label, the section itself was essentially mislabeled. Family and Social History are particularly important as they represent the most basic form of genetic information and can impact guideline-based clinical care decisions that rely on risk stratification\textsuperscript{18}. The selected evaluation criteria account for these errors by measuring the amount of likely relevant content (i.e., the number of UMLS concepts) captured. The result of the findings presented is that the identification of Family and Social History information had very poor precision when using a method that looks for section header labels. Additionally, the HMM performed consistently from training to testing sets compared to TBM, indicating that it may be a more portable method for note section detection across different templates and note types.

When considering the NLP tasks that are downstream of note section identification, the HMM results are particularly promising. Two of the three note section types for which the HMM performed the worst are for data that are relatively low yield in clinical notes. Both Medications and Labs and Imaging data within notes are often copied from structured sources or dedicated reports. These data are likely more useful in the structured form from which they were derived than extracted from clinical notes.

Despite these findings, the methods presented have significant room for improvement, and analyzing common errors can inform future model development. Improved detection of candidate section transitions would improve the accuracy of the model by decreasing the number of state change decision points. This could be accomplished with additional text preprocessing to, for example, reformat the notes so that narrative text that is broken up into multiple lines intra-paragraph and intra-sentence is considered as a single block inside which transitioning of states cannot occur. In addition, staging the identification of sections by starting with an HMM that only considers the three categories of “Subjective”, “Objective”, and “Assessment and Plan,” and then further processing of each of those into component sections may improve performance.

The promising results of this study suggest that clinical note sections can be identified using an HMM based on UMLS semantic types associated with MetaMap identified UMLS CUIs. The relatively standard organization of clinical notes and the expected information contained within each section is both the motivation for identifying clinical text sections and the reason why these methods worked. It is likely that these methods would be portable across EHR systems, as preliminary testing on an additional corpus of publicly available notes from MTSamples.com demonstrates better performance than reported here. The ability to identify note sections across EHR systems and medical specialties using minimal training data in this way would enable the development of tools to aggregate clinical data about an individual patient across health systems. For example, the aggregation of sparsely recorded information like Family and Social History could be collected and synthesized into a single best summary for a patient. This information may be more likely to be recorded by a primary care physician but could then be made available to physicians during all of a patient’s clinical encounters. In addition, the note section contains contextual information about the meaning of the underlying information. Most importantly, distinguishing between Medical History and Assessment and Plan is needed in order to assemble a chronology of a patient’s clinical information. For example, a medication that appears as part of a standard Medications section is generally confirmed as active therapy, while a medication that appears in the Assessment and Plan could go unfilled.

All code written for this paper is available through GitHub (http://github.com/aeisman/NoteSectDetect).

Conclusion

The hidden Markov model and results presented are the first step in the development of a generalizable method for the identification of sections based on content in clinical notes. The findings demonstrate that semantic types associated with MetaMap identified UMLS CUIs differ among clinical note sections, and the probability of observing consecutive content within notes can be used to determine note sections. Additional work is required to improve the performance of these models and test this method on a variety of note types across different healthcare settings.

Acknowledgments

The research reported in this publication was supported in part by the National Institutes of Health grants F30LM013320, U54GM115677, and R25MH116440. The contents are solely the responsibility of the authors and do not necessarily represent the official views of the National Institutes of Health.

426
References

1. Everson J, Rubin JC, Friedman CP. Reconsidering hospital EHR adoption at the dawn of HITECH: implications of the reported 9% adoption of a “basic” EHR. J Am Med Inform Assoc [Internet]. 2020 Aug 1;27(8):1198–205. Available from: http://dx.doi.org/10.1093/jamia/ocaa090
Abstract The wide availability of near infrared light sources in interventional medical imaging stacks enables non-invasive quantification of perfusion by using fluorescent dyes, typically Indocyanine Green (ICG). Due to their often leaky and chaotic vasculatures, intravenously administered ICG perfuses through cancerous tissues differently. We investigate here how a few characteristic values derived from the time series of fluorescence can be used in simple machine learning algorithms to distinguish benign lesions from cancers. These features capture the initial uptake of ICG in the colon, its peak fluorescence, and its early wash-out. By using simple, explainable algorithms we demonstrate, in clinical cases, that sensitivity (specificity) rates of over 95% (95%) for cancer classification can be achieved.

1 Introduction

The wide availability of near infrared (NIR) light sources in interventional medical imaging stacks enables non-invasive quantification of perfusion by using fluorescent dyes, typically Indocyanine Green (ICG). As a result, fluorescence-guided surgery, where ICG is utilized e.g. for lymph node mapping and identification of solid tumors, has become common practice. Interrogation of tissue and vasculature using ICG has also found application in the assessment of anastomotic complications and the prediction of suspected cancerous tissue. The prediction of cancerous tissue is not sufficiently useful currently due to false positive rates and problems of interpretation.

It has been reported that ICG accumulates in cancers, which can offer a means of differentiation between healthy and malignant tissues, however grossly obvious differences that are detectable by the human eye can take hours to appear. In contrast, attempts to leverage the short-time differences in fluorescence behaviour, namely how long it takes for fluorescence to reach its peak and at what rate it subsequently decays, which requires on the order of only 10 minutes of observation has been shown to be feasible by exploiting the bio-physics of perfusion processes.

Since we are primarily interested in perfusion dynamics, i.e. the evolution of fluorescence in specific tissue over time, the task is one of surgical video classification. This task cannot be readily translated to an image classification problem for several reasons, precluding the use of existing methods. Although containing many individual images, each surgical video represents only a few sufficiently distinct images and results in significant duplication when used as a training set. Indeed, most of the frames of the NIR video are near duplicates when compared e.g. by using perceptual hashing.

In our approach, we refine the bio-physical two-compartment model of ICG dynamics for the purposes of simplifying computation and improving interpretation of results. Our contributions differ from previous work in the following important aspects: (1) The number of features used is greatly reduced from previous models and their significance is obvious. (2) Feature values have clear meanings such as the slope of the initial uptake of ICG, the time it takes to reach peak perfusion, and the rate of decay of ICG once the peak is reached (this last feature is – to the best of our knowledge – absent from the previous literature). These features can be estimated from visual inspection of the fluorescence time-series addressing interpretability of results. (3) Inter-patient variation is taken into account by setting each feature in relation to a healthy reference region in the same patient.

Related Work. Cancer angiogenesis has been recognized to lead to abnormal vasculature, characterized by higher internal pressure, wide inter-endothelial junctions, high number of trans-endothelial channels and a discontinuous basement membrane, in comparison with normal, healthy tissue. Cancer detection methods based on analysis of multispectral videos have become popular with the advent of advanced intra-operative fluorescence imaging systems.
and progress in image processing. We consider them in three broad categories: (1) observation by the surgeon of perfusion in the colon (2) image classification using Deep Learning (DL), and (3) quantitative analysis of bio-physical models based on perfusion patterns.

Observation of perfusion patterns by surgeons is common practice for anastomoses. However, the changes in the perfusion needed to discriminate cancer are not easily detected by visual observation. Considerable time needs to elapse, in the order of days, for easy visual discrimination, and this still remains prone to error, rendering application during surgery problematic.

Deep learning, a sub-discipline of artificial intelligence, has been applied for cancer screening using images from mammography and histology. Deep learning methods typically require enormous numbers of labelled training examples for each class to estimate model parameters. Cancer screening is observed to require on the order of tens of thousands of mammograms with known pathology. A further drawback of deep learning, which we wish to avoid in the presented approach, is the lack of explainability inherent to the complex path between inputs (i.e. images, videos, annotations, etc) and classifications.

Mathematical models describing the bio-physics of perfusion provide an alternative method of fluorescence analysis. Perfusion quantification based on estimating time-series of ICG fluorescence intensities, and then extracting a number of so called time-to-peak (TTP) features directly from the time-series is well represented in the surgical literature. TTP features have direct physical meaning and have been successfully applied to perfusion quantification on animal, and for predicting anastomotic complications after laparoscopic colorectal surgery. By exploiting models of the underlying perfusion process, bio-physical methods require vastly less annotated image data than deep learning methods and use more clinically recognizable features.

The premise of bio-physical modeling for perfusion imaging is that differences in ICG inflow during a relatively short timeframe of minutes after injection known as the wash-in phase, and ICG outflow during the beginning of the venous outflow, termed the wash-out phase, can serve as a marker for tissue distinction. Relevant to cancer, ICG inflow over a wash-in phase could disclose valuable discriminative signatures. It is also noted that most malignant tumors are characterized by “increased interstitial pressure from leaky vessels” and “relative absence of intra-tumoral lymphatic vessels”, suggesting that fluid “could pool” in the malignant tumor over time, causing the retention of the tracer, leading to differences of ICG outflow during wash-out. These behaviors are different between cancerous tissue and benign tumors and healthy tissue.

The bio-physical model proposed in generalize models proposed in by modeling ICG intensity time-series as a response of a generic second-order linear system with exponential input, which is a sum of one real and two complex exponentials, to allow for oscillating behaviors observed in ICG time-series estimated from videos of human tissue perfusion. The coefficients and exponents of these exponential terms form a set of features.

To date there are no commercially available systems to aid in the characterization (benign vs malignant) of colorectal lesions. Endoscopic adjuncts such as narrow band imaging (NBI) have been shown to help differentiate lesions but only in the hands of experts and require considerable user interpretation. Combining NBI with AI interpretation has been reported however this was to characterize small polyps of the colorectum as either hyperplastic (non-neoplastic) or adenomatous (neoplastic) polyps and was not used to identify malignancy. Large multicentre studies consistently demonstrate the limited diagnostic accuracy of conventional endoscopy, most especially for early malignancies where accuracy may be as low as 39%.

2 Dataset

Following full Institutional Approval (1/378/2092), colorectal tumors from 16 patients with known rectal lesions undergoing an examination under anaesthesia for either diagnostic or surgical planning purposes were imaged transanally for up to 30 continuous minutes following Indocyanine Green (ICG) administration (0.25mg/kg i.v) using a Pinpoint (Novadaq, Stryker) near infrared (NIR) imaging system. The resulting video is multispectral: the NIR image is stored separately from the visible light image, see Figure 1 for an example. The frame rate is 29.95 FPS.

Representative tissue samples were obtained through biopsy or excision in keeping with standard clinical practice. Subsequent histopathological analysis of specimens, along with specialist colorectal clinician input, facilitated retro-
Figure 1: Example video frame: the upper panel shows the visible light, the lower panel shows the NIR image.

Figure 2: Tracking, i.e. compensation of motion and deformation in the videos, is challenging: the frames on the left illustrate regions having drifted from their original locations, likely due to reflections and lack of textures. The frames on the right show an example of successful tracking. The lesion on the left turned out to be benign, whereas the one on the right is malignant.

spective annotation of tissue regions within each video (“cancer”, “benign” or “healthy”), resulting in 5 patients with cancerous regions, and 11 patients with benign lesions. Quantitative fluorescence plots over time for regions of interest based on video annotations were subsequently created for analysis using a bespoke tracker: movement of patient and camera as well as deformations, reflections, etc have to be compensated for in order to collect fluorescence in one region of tissue (as opposed to one location within the image) over time, see Figure 2 for an illustration. The tracker will be the subject of subsequent publications; we stress however that its performance was assessed for each video and regions were discarded if tracking was not deemed precise enough; for the present dataset, 3 out of 180 tracked ROIs were discarded because of imprecise tracking.

Videos were considered suitable for inclusion if stable mucosal views of both tumor and adjacent healthy/control tissue were successfully tracked for the duration of ICG ingress, peak and at least 8 minutes of ICG outflow.

3 Methods
3.1 Video Processing

The underlying assumption to all what follows is that fluorescence intensity is proportional to the ICG concentration in the tissue, which has been shown in the literature for the concentrations commonly encountered. Following this assumption, the brightness in the NIR image can be interpreted as a relative measurement of ICG concentration.

From each available video \( p \), between 3 and 8 fluorescence time-series are extracted in the following steps: (i) a clinical expert identifies areas of healthy tissue and areas of concern, (ii) in each area, depending on its size, between 1 and 4 rectangular regions of interest (ROIs) are drawn. Typically, 4 ROIs are drawn on the area of concern, and 4 ROIs are drawn on the healthy tissue. (iii) The ROIs are tracked through the video for as long as they are visible. This time interval has to start before ICG arrives in the tissue under observation and has to include at least the peak fluorescence and 8 minutes thereafter. The tracking is performed in the visible light video, since the only changes there are due to movement, deformation, occlusions etc, and ICG is not visible there. (iv) For each ROI \( r \) and each included time step

\[ \text{by that we mean the absolute concentration is not available, however if a region at time } t_1 \text{ appears twice as bright as the same region at } t_0, \text{ we may conclude that the ICG concentration in that region has doubled between } t_0 \text{ and } t_1. \]
Figure 3: Example profiles, with circles and vertical dashed lines indicating the detected peaks and latencies, solid lines are smoothed, dashed lines indicate the original, noisy data. While in some cases, the collected data is pretty smooth already (left), in some cases it is so noisy that even discerning a clear peak is hard; yet other cases have initial peaks which are likely due to camera movement and/or adjustments of internal camera parameters. Incidentally, the orange curves correspond to healthy reference regions, whereas the blue curve in the left panel corresponds to a cancer, and the remaining curves to benign lesions.

The result is a collection of time series $I_{p,r}(t)$, where $p$ ranges from 1 to 16 (the number of patients in our dataset), $r$ ranges from 1 to the number of ROIs tracked in the specific patient (typically 8), and $t$ ranges from 0 seconds to up to 30 minutes, depending on how long ROI $r$ in patient $p$ was in view.

3.2 Peak and Latency Estimation

The resulting time series can be quite noisy, see Figure 3 for a few examples. It is hence necessary to smooth the data before proceeding with the detection of peaks and the estimation of the latency. By “peak” we here mean when the fluorescence in each region reaches its highest value, and by “latency” we mean the point in time when fluorescence starts rising beyond the initial background fluorescence.

The smoothing is done using a Savitzky-Golay filter\textsuperscript{29} of order 5 with window length 99. These values were chosen by manually annotating peak and latency in a small subset of our data and optimizing with respect to average estimation error. The peak detection is done using its SciPy\textsuperscript{30} implementation, whereas the latencies $L_{p,r}$ are detected using a custom detector considering the smoothed derivative of the curves and finding its first “robust” zero crossing.

Since medical imaging devices are typically not designed to be measurement instruments, but rather to present a readily interpretable image to the user\textsuperscript{31,32}, it is clear that absolute NIR brightness is not meaningful: it depends on the distance of the ROI to the camera, the distance from the ROI to the center of the frame, and most importantly on the camera parameters, which are typically chosen to optimize contrast and overall visual representation\textsuperscript{31,32}. To counteract this, every ROI’s brightness is normalized by its peak brightness, so that the peak value equals 1 for each of them.

3.3 Feature Design

The features are chosen in two steps: First, the following characteristic numbers are chosen for each ROI individually:
The time to peak (TTP\(_{p,r}\)) is simply the time difference between when the peak is reached and the latency \(L_{p,r}\).

The upslope \((U_{p,r})\) is computed as \(U_{p,r} = \frac{1-I_{p,r}(L_{p,r})}{TTP_{p,r}}\), i.e. it is the average slope between initial ICG arrival and the peak.

The downslopes \(DX\) are the average downslopes between the peak and \(X\) seconds further, so \(DX_{p,r} = \frac{1-I_{p,r}(L_{p,r}+TTP_{p,r}+X)}{X}\). We included \(DX\) for \(X \in \{20, 50, 80, 100, 160, 250, 400\}\) seconds in our initial feature set, but only the most discriminative values of \(X\) will be used in the final classifier.

Since we noticed that estimating the downslope based on a single time point might lead to very noisy estimates, see e.g. the middle panel of Figure 3, a more robust feature was introduced by estimating the downslopes of a window around \(L_{p,r}+TTP_{p,r}+X\) seconds and taking the median of those downslopes:

\[ DX_{p,r,avg} = \text{median}_{x \in \text{Window}} D(X + x)_{p,r}, \quad \text{Window} = \{-15, -13, \ldots, 13, 15\} \text{seconds} \]

The time ratio \(TR\) is included as well. It is the ratio between \(TTP_{p,r}\) and when \(I_{p,r}(t)\) reaches half the peak values, see Figure 4.

While \(U\) and TTP relate to the uptake of ICG, the \(DX\) relate to the decay of ICG fluorescence; \(TR\) is a measure of the temporal inhomogeneity of the initial uptake.

Once this first set of features is obtained, inter-patient variation can be addressed by relating the features of each ROI to a healthy reference value in the same patient. In particular, for a feature \(F\), we define its patient-normalized value \(f\) to be

\[ f_{p,r} = \frac{F_{p,r} - \text{median}_{\text{is "healthy"}} F_{p,s}}{\text{median}_{\text{is "healthy"}} F_{p,s}}, \]

in other words, we are choosing the median value (as an average that is robust to outliers) of a feature across a patient’s healthy tissue as a reference value, and define each feature as its percentage difference to that reference value.

This puts into numerical terms the intuition that “benign” growths would have similar perfusion patterns as “healthy” tissue, whereas “cancer” tissue should deviate from that. As an illustration, consider Figure 5: the normalized downslope feature concentrates around 0 for benign regions, whereas it concentrates away from 0 for the cancerous regions; for other features, e.g. the time-to-peak TTP, that appears to not be the case, hence it is necessary to consider all features in the feature selection process.

### 3.4 Classification Algorithms and Feature Selection

While the data acquisition required addressing technical challenges in the realm of computer vision, and the feature extraction and design required domain insights, the classification itself is now a rather standard binary classification problem, with feature selection as a subproblem: Given the features of a region for which the pathology is unknown,
we want to assign the label “benign” or “cancer” to it. Additionally, we want to explore which small subset of features achieves the best performance.

Considering the scarcity of data and the justified desire for simple, interpretable algorithms in the context of digital health, we restrict ourselves to a subset of available machine learning (ML) algorithms. In particular, we exclude neural networks from consideration, and evaluate support vector machines (SVMs), naive Bayesian classifiers, generalized additive models (GAMs), decision trees, nearest neighbors, and logistic regression. These standard algorithms are well-described in the literature and have simple to use, efficient Python implementations.

Furthermore, we restrict the number of finally used features to two, for: (a) this enables us to draw decision boundaries, see Figure 6 (b) with a dataset as small as ours, there is no justification for using high-dimensional feature spaces.

3.5 Feature Selection

We selected the best pair of features in terms of achieved sensitivity, specificity, and accuracy by a two-step procedure:

- In a first step, decision trees and SVMs using the full set of 34 features defined above were trained, and recursive feature elimination (RFE) was performed to arrive at a much smaller set of best-performing features. Those were \{d_{250\text{avg}}, d_{400\text{avg}}, tr, TR\}.

- In the second step, an exhaustive search over pairs of features was performed, by training classifiers for the 6 possible pairs of two chosen from the 4 features identified in the previous step.

In all cases, performance was assessed by a k-fold cross-validation strategy where a portion of the training data is reserved and used to test the model after hyperparameter optimization; this is detailed in Section 4.

Care also needs to be taken to prevent leakage from training into testing sets: Since ROIs from the same patient often have similar characteristics, splitting into testing and training sets was done on the patient level, i.e. all of a patient’s ROIs will always be either in the training or the testing set, but never in both. Experiments were run with 20 random splits of training and test data sets and the mean of the performance metrics are reported for sensitivity, specificity and accuracy respectively.

4 Results

Performance of the interpretable machine learning models was assessed by first selecting 12 patients to be the training set, and setting the remaining 4 patients aside as the test set. Each ML model was then tuned to the training set in an inner 5-fold cross validation procedure (i.e. the remaining 12 patients were split into 5 folds, 4 of which were used to train the classifiers, and the last fold was used to assess performance), and evaluated by its performance on the test set. This process was repeated 20 times (there would be 1820 ways to split 16 into groups of 12 and 4), and the average of each metric was reported as the classifier’s performance. Table 1 shows the best model performances with all pairs of
Table 1: For all feature pairs, best performing algorithms for 3 metrics and the average metric (Accuracy + Specificity + Sensitivity)/3. NB = Naive Bayes, NN = Nearest Neighbors, DT = Decision Tree, GAM = Logistic Generalized Additive Model, LogReg = Logistic Regression.

<table>
<thead>
<tr>
<th>Features</th>
<th>Best Average</th>
<th>Best Accuracy</th>
<th>Best Sensitivity</th>
<th>Best Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>$d_{250_{avg}}$, $d_{400_{avg}}$</td>
<td>NB 0.95 (0.10)</td>
<td>NB 0.95</td>
<td>GAM 0.96 (0.13)</td>
<td>NN 0.96 (0.09)</td>
</tr>
<tr>
<td>$d_{250_{avg}}$, tr</td>
<td>NB 0.93 (0.12)</td>
<td>NB 0.93</td>
<td>LogReg 0.95 (0.22)</td>
<td>NN 0.96 (0.09)</td>
</tr>
<tr>
<td>$d_{250_{avg}}$, TR</td>
<td>NN 0.90 (0.09)</td>
<td>NN 0.93</td>
<td>LogReg 0.95 (0.22)</td>
<td>NN 0.96 (0.09)</td>
</tr>
<tr>
<td>$d_{400_{avg}}$, tr</td>
<td>NN 0.89 (0.09)</td>
<td>NN 0.93</td>
<td>LogReg 0.95 (0.22)</td>
<td>NN 0.96 (0.09)</td>
</tr>
<tr>
<td>$d_{400_{avg}}$, TR</td>
<td>NN 0.87 (0.09)</td>
<td>NN 0.92</td>
<td>LogReg 0.95 (0.22)</td>
<td>NN 0.96 (0.09)</td>
</tr>
<tr>
<td>TR, tr</td>
<td>DT 0.81 (0.16)</td>
<td>NN 0.87</td>
<td>LogReg 0.91 (0.25)</td>
<td>NN 0.92 (0.15)</td>
</tr>
</tbody>
</table>

Figure 6: The decision boundary of a naive Bayes classifier based on the feature combination $d_{250_{avg}}$ and $d_{400_{avg}}$ and a decision tree based on $d_{250_{avg}}$ and tr. Each plot can be interpreted as follows: For a given ICG brightness curve, compute or estimate the features and locate the point in the plot. The deeper the shade of red, the higher the classifier’s confidence in “cancer”, and the deeper the shade of blue, the higher the confidence in “benign”. The dots denote the data used to train the classifiers (a random selection of 12 patients).

The classifier with the best average of all considered performance metrics is a naive Bayes classifier using the features $d_{250_{avg}}$ and $d_{400_{avg}}$, which achieves an Accuracy/Sensitivity/Specificity of 0.953/0.950/0.958. To illustrate the resulting classifiers and their interpretable nature, we computed the decision boundaries for this best classifier, and, for comparison, a decision tree using the features $d_{250_{avg}}$ and TR, see Figure 6.

Conclusions
We stress a few important points:

1. While the process of preparing the data and training the algorithms is complex, the resulting algorithms them-
selves are so simple that the predicted pathology can be computed in one step from a simple tabulation of a
ROI’s brightness over the course of around 7 minutes.

2. The more traditional features TTP and $U$ do not appear in our selection of most discriminative features, however
two of the novel downslope features derived from $D_{250}$ and $D_{400}$ do appear. TTP is of course still necessary
to compute $DX$ (recall that it is the average slope between TTP and TTP + X seconds).

3. The training set is very limited, hence while numbers in excess of 90% are exciting and encouraging, we caution
that significantly more videos will be necessary to make assessments with any confidence.

To summarize: We proposed to characterize dynamic perfusion patterns by a number of easily interpretable features
(e.g. the slope of the initial uptake of ICG, the time it takes to reach peak perfusion, and the rate of decay of ICG
once the peak is reached). These features can be estimated merely by visual inspection of the fluorescence time-series
addressing interpretability of results, and computed by hand or in a spreadsheet. For validation of the proposed features
we implemented an experimental framework combining a tracking algorithm, basic peak and latency estimation
algorithms (which could be replaced by visual inspection of the selected data) and classification algorithms from stan-
dard machine learning open source tools. Experiments on a corpus of 16 suspected colorectal cancer multispectral
endoscopic videos (5 confirmed cancer/11 confirmed benign lesion) captured during surgical procedures, showed that
using only 4 of the proposed features is discriminant for benign and malignant tissues; reducing the number of features
to 2 is seen to decrease performance, but not dramatically, while lending itself to a simple visualization of the decision
boundaries between benign and malignant tissue.

The method of tissue characterization described here has many potential uses. Immediate uses include guiding the
operator to the tissue locations most likely to yield a diagnostic biopsy as well as the identification of tumor margins.
Following further refinement of the technique, confident characterization of colorectal lesions at the time of initial
encounter would accelerate the patient towards definitive care, whether that is proceeding with local excision (e.g.
Trans-anal minimally invasive surgery (TAMIS) in the case of benign or early stage malignancy), or more extensive
surgical resection.

References

1. Luigi Boni, Giulia David, Gianlorenzo Dionigi, Stefano Rausei, Elisa Cassinotti, and Abe Fingerhut. Indocya-
nine green-enhanced fluorescence to assess bowel perfusion during laparoscopic colorectal resection. *Surgical

2. Yeon-Ju Huh, Hyuk-Joon Lee, Tae-Han Kim, Yun-suck Choi, Ji-Ho Park, Young-Gil Son, Yun-Suhk Suh, Seong-
Ho Kong, and Han-Kwang Yang. Efficacy of assessing intraoperative bowel perfusion with near-infrared camera
483, 2019.

3. Boudewijn E Schaafsma, J Sven D Mieog, Merlijn Hutteman, Joost R Van der Vorst, Peter JK Kuppen,
Clemens WGM Lówik, John V Frangioni, Cornelis JH Van de Velde, and Alexander L Vahrmeijer. The clinical

4. Gyung Mo Son, Myeong Sook Kwon, Yoonhong Kim, Jisu Kim, Seung Hwa Kim, and Jung Woo Lee. Quan-
titative analysis of colon perfusion pattern using indocyanine green (icg) angiography in laparoscopic colorectal

5. Faïçal Selka, Vincent Agnus, Stéphane Nicolau, Abdel Bessaid, Luc Soler, Jacques Marescaux, and Michele
Diana. Fluorescence-based enhanced reality for colorectal endoscopic surgery. In *International Workshop on

6. Jagadeesan Jayender, Eva Gombos, Sona Chikarmane, Donnette Dabydeen, Ferenc A Jolesz, and Kirby G Vos-


15. DRC James, Frédéric Ris, TM Yeung, R Kraus, NC Buchs, NJ Mortensen, and RJ Hompes. Fluorescence angiography in laparoscopic low rectal and anorectal anastomoses with pinpoint perfusion imaging-a critical appraisal with specific focus on leak risk reduction. *Colorectal Disease*, 17:16–21, 2015.


23. Michael Gurfinkel, Alan B Thompson, William Ralston, Tamara L Troy, Ana L Moore, Thomas A Moore, J De- 
vens Gust, Derreck Tatman, Jeffery S Reynolds, Bruce Muggenburg, Kristin Nikula, Ravindra Pandey, Ralf H. 
Mayer, Daniel J. Hawrysz, and Eva M. Sevick-Muraca. Pharmacokinetics of ICG and HPPH-car for the detection 
of normal and tumor tissue using fluorescence, near-infrared reflectance imaging: A case study. Photochemistry 

24. Colin J Rees, Praveen T Rajasekhar, Ana Wilson, Helen Close, Matthew D Rutter, Brian P Saunders, James E 
East, Rebecca Maier, Morgan Moorghen, Usman Muhammad, Helen Hancock, Anthoor Jayaprakash, Chris Mac- 
Donald, Arvind Ramadas, Anjan Dhar, and James M Mason. Narrow band imaging optical diagnosis of small 
colorectal polyps in routine clinical practice: the detect inspect characterise resect and discard 2 (DISCARD 2) 

25. Yuichi Mori, Shin-ei Kudo, Masashi Misawa, and Kensaku Mori. Simultaneous detection and characterization of 

26. Jasper LA Vleugels, Lianne Koens, Marcel GW Dijkstra, Britt Houwen, Yark Hazewinkel, Paul Fockens, and 
Evelien Dekker. Suboptimal endoscopic cancer recognition in colorectal lesions in a national bowel screening 

27. Yara Backes, Matthijs P Schwartz, Frank ter Borg, Frank H J Wolfhagen, John N Groen, Wouter H de Vos tot 
Nederveen Cappel, Jeroen van Bergeijk, Joost M J Geesing, Bernhard W M Spanier, Paul Didden, Frank P 
Vleggaar, Miangela M Lacle, Sjoerd G Elias, and Leon M G Moons. Multicentre prospective evaluation of 
real-time optical diagnosis of t1 colorectal cancer in large non-pedunculated colorectal polyps using narrow band 

28. R C Benson and H A Kues. Fluorescence properties of indocyanine green as related to angiography. Physics in 

29. Abraham Savitzky and M. J. E. Golay. Smoothing and differentiation of data by simplified least squares proce- 

30. Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni 
Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua 
Carey, Ilhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, 
Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, 
Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in 

31. J. Fengler. Near-infrared fluorescence laparoscopy - technical description of PINPOINT® a novel and commer- 
cially available system. Colorectal Disease, 17:3–6, 2015.

32. Jeffrey Dalli, Niall Hardy, Pol G. Mac Aonghusa, Jonathan P. Epperlein, Padraig Cantillon Murphy, and Ronan A. 
Cahill. Challenges in the interpretation of colorectal indocyanine green fluorescence angiography — a video 
vignette. Colorectal Disease.

33. Jerome Friedman, Trevor Hastie, and Robert Tibshirani. The elements of statistical learning, volume 1. Springer 

34. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, 
V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: 


Launching into clinical space with medspaCy: a new clinical text processing toolkit in Python

Hannah Eyre, MS1,2, Alec B Chapman, MS1,2, Kelly S Peterson, MS2,3, Jianlin Shi, MD, PhD2, Patrick R Alba, MS1,2, Makoto M Jones, MD1,2, Tamára L Box, PhD2, Scott L DuVall, PhD1,2, Olga V Patterson, PhD1,2
1VA Salt Lake City Health Care System; 2University of Utah, Salt Lake City, UT, USA; 3Veterans Health Administration Office of Analytics and Performance Integration

Abstract Despite impressive success of machine learning algorithms in clinical natural language processing (cNLP), rule-based approaches still have a prominent role. In this paper, we introduce medspaCy, an extensible, open-source cNLP library based on spaCy framework that allows flexible integration of rule-based and machine learning-based algorithms adapted to clinical text. MedspaCy includes a variety of components that meet common cNLP needs such as context analysis and mapping to standard terminologies. By utilizing spaCy’s clear and easy-to-use conventions, medspaCy enables development of custom pipelines that integrate easily with other spaCy-based modules. Our toolkit includes several core components and facilitates rapid development of pipelines for clinical text.

Introduction

Retrospective clinical research often relies on data extracted from electronic medical record (EMR) systems using natural language processing (NLP) adapted for clinical language. Despite a wide range of existing solutions, code reuse is challenging because new system development projects face the labor-intensive task of connecting isolated modules into cohesive cNLP pipelines. Over the years, the need for reuse and reproducibility has been met with a number of frameworks and toolkits.1 Java-based general architectures, such as UIMA2 and GATE3, have provided a strong foundation for a number of highly successful cNLP toolkits and general purpose reusable systems such as cTAKES, CLAMP, Leo, and HITE4–7. As a platform-independent and straightforward programming language, Java has been the primary environment for most cNLP applications of the last 20 years even prompting re-implementation of tools previously written in other programming languages to improve accessibility8,9.

While Python programming language is celebrating its 30-year anniversary, in the last few years its usage has exploded as it had become one of the most popular languages of data science.10–14 The wider NLP, machine learning, and data science community’s shift towards Python is fueled by its capability of interactive environments to explore data and experiment with approaches while avoiding a common cycle of compiling code and reloading data15. The machine learning ecosystem in Python is vibrant and growing, including state-of-the-art training methods and deep learning models.

Until recently Natural Language Toolkit (NLTK) has been the leading platform for general text processing in Python.16 Actively supported by a dedicated team since 2000, NLTK is comprised of a number highly functional NLP libraries for rule-based and machine learning-based text analysis. While NLTK has enabled hundreds of successful research and educational projects, its approach to text processing as lists of strings and lack of a unifying architecture have hindered its ability to scale to large datasets.

In contrast, spaCy+ provides a robust architecture for building and sharing custom, high-performance NLP pipelines by taking an object-oriented view of text. It is non-destructive, supports seamless integration of statistical and machine-learning methods with rule-based NLP, and allows for the creation of custom components for specialized tasks. Powered by the strength of Cython, an optimizing static compiler for Python that generates very efficient C or C++ code, spaCy allows achieving exceptional speed of performance. Similar to UIMA approach in Java, spaCy provides a framework for modular plug-and-play construction of custom NLP pipelines. Since its inception in 2015, spaCy universe has gained a robust, highly-active, and growing community of contributors of open source modules, state-of-the-art models, and end-to-end systems developed with or for the framework.

Recognizing the need to adapt processing to different domains, several models and toolkits have been introduced targeting specialized text, among them scispaCy that handles scientific and biomedical sublanguages17. While scispaCy includes custom tokenization rules, biomedical concept identification, and a variety of pre-trained dependency parsing

*https://spacy.io/*
and named entity recognition (NER) models, the differences between clinical and biomedical language hamper its ability to achieve optimal performance on clinical narrative. To confront the particular challenges posed by medical text, a team at Virginia Commonwealth University created MedaCy, a medical text mining framework built over spaCy to facilitate the engineering, training and application of machine learning models for medical information extraction\textsuperscript{18}. Neither scispaCy nor MedaCy provide adequate support for rule-based system development beyond the functionality natively provided by spaCy. Despite a wide range of machine learning applications, rule-based algorithms still comprise a large proportion of successful academic and commercial cNLP projects\textsuperscript{19, 20}.

To bridge the gap between the existing functionality and the needs of the clinical research community for easy-to-use clinical text processing that combines statistical and symbolic methods, a team of enthusiastic cNLP practitioners at the Veterans Health Administration and University of Utah developed a spaCy-based library of core components targeting medical text called medspaCy\textsuperscript{†}.

**Methods**

**Framework Overview**

MedspaCy is designed to utilize spaCy’s application programming interface (API) while only minimally expanding upon it where needed. This minimal expansion of spaCy’s API ensures each component is usable standalone or as one part of a larger spaCy pipeline that may incorporate components from other sources. Figure 1 illustrates the interconnection of medspaCy components into an end-to-end system through a common API. Once the toolkit reached acceptable maturity, we made it available through PyPI with the command `pip install medspacy`.

![Figure 1: Overview of medspaCy architecture.](image)

All information generated by components is directly accessible through the use of spaCy’s attribute extension functionality. For example, a default spaCy attribute of a named entity is accessible using `ent.like_num` attribute, a medspaCy attribute utilizes the custom extensions such as `ent..is_negated`. Attribute extensions are registered at the `Doc`, `Span`, or `Token` levels, depending on the component to create a variety of useful tools for analyzing data after medspaCy processing is complete.

The components included in medspaCy offer both a default initialization and a large number of optional customization

\textsuperscript{†}https://github.com/medspacy/medspacy
parameters. This allows components to be quick to set up, learn, and use to develop prototypes while also being fully customizable for more specific needs. Resources, such as curated rule sets or knowledge bases, may be shared with the community and utilized across projects. Figure 2 is an illustration of an example pipeline using medspaCy. Since the spaCy API allows for modular components, there are many possible pipelines where users could insert alternative components, such as scispaCy AbbreviationDetector or other spaCy-compatible rule-based or machine learning-based modules.

**Figure 2:** Example of a text processing pipeline that utilizes medspaCy and other spaCy-based components.

**MedspaCy Components**

MedspaCy has a growing list of integrated components that allow building end-to-end systems for clinical text processing.

**Tokenization**

The primary advantage of spaCy’s approach to tokenization is that it is non-destructive, which preserves all whitespace and punctuation information enabling complete reconstruction of the original text. The major drawback of the default spaCy tokenizer for clinical text processing is that it is not trained on clinical text. Additionally, it has a variety of rules designed to handle text sourced online, including many rules that mitigate excess tokenization of URLs. These rules prevent splitting sequences of alphanumeric characters and punctuation into multiple tokens. However, URLs are relatively uncommon in clinical text but typos and using punctuation to delineate document structures are common. The tokenizer included in medspaCy implemented custom rules to handle punctuation and inconsistent use of whitespace that are common in clinical notes.

**Sentence Detection**

Due to the limited purpose of clinical language to document and succinctly communicate information about a specific patient, sentences in clinical text are characterized by telegraphic grammar with omitted subjects and the propensity to use long lists for present or absent conditions and prescribed medications\(^ {21} \).

Sentence segmentation in clinical language is hindered by frequent use of templated and tabular text, abundance of concept-value pair statements, and inconsistent use of punctuation. Since the text is meant to be read within a specific user interface of the EMR, space, tab, and new line characters are utilized to indicate sentence or section boundaries as well as for visual indication of grouped text, such as in case of templates and tabular text. Syntactically significant whitespace complicates the traditional approach to text processing that ignores existence and length of whitespace in text. To correctly identify the context of target concepts from clinical text, a clinical-domain sentence boundary detector is essential, because many downstream components work at the sentence level. Commonly used sentence detectors trained on non-clinical text typically do not perform well on clinical text data.\(^ {22} \) To make medspaCy more capable of performing cNLP, we adopted a high-performance, rule-based sentence detector for clinical text – PyRuSH, which is the Python version of Rule-based sentence Segementer using Hashing (RuSH)\(^ {22} \). Relying on rules for tokenization allows easy adaptation of the tokenizer to a new setting without extensive retraining.
Section Detection

Clinical documentation has a well established data format for different document types. Logical Observation Identifiers Names and Codes (LOINC) Document Ontology outlines hundreds of document types that vary by subject domains, role of the author, setting, type of service described, and document kind\textsuperscript{23}. At each intersection of these axes, a clinical document has a specific structure that is either learned through apprenticeship (medical students learning from practice) or specifically prescribed by guidelines\textsuperscript{24,25}. The same statement in different sections of the same or different documents might have very different meanings. For example, a document type called Progress Note has Subjective, Objective, Assessment, Plan sections, and Procedure Report has Findings and Impressions sections. Depending on which section it is mentioned in, a clinical condition may be experienced by a patient in the past, may be happening at the time when the note is written, or the patient may be at risk for experiencing it in the future. These sections may or may not be explicitly labeled with a variety of section headers fully spelled out or abbreviated.

Breaking documents into relevant sections is often a core part of cNLP, especially as documents grow in length and complexity. medspaCy includes an implementation of clinical section detection based on rule-based matching of the section titles with default rules adapted from SecTag\textsuperscript{26} and expanded through practice. The sectionizer contains several other features beyond simple section title identification. If desired, the sectionizer can interact with the attributes of entities, such as changing the temporality attribute of all entities in the “Past Medical History” section. The medspaCy sectionizer is also capable of creating hierarchies of sections and subsections within the document and preserving them in a traversable tree structure.

Concept Extraction

Clinical concept extraction through concept mention detection and concept encoding is one of the most prominent tasks of clinical text processing.\textsuperscript{19} Similar to NER, concept extraction employs either manual rules or trained statistical models to identify specific spans of text from a clinical document and labeling them with pre-defined clinical concepts. Several utilities are included in medspaCy for targeted concept extraction which extend existing spaCy rule-based functionality. Rules defining concepts specify the text pattern to be matched, semantic category, and additional metadata about a concept. In addition to integrating existing pattern-matching functionality provided by spaCy, medspaCy incorporates additional regular expression features, such as accepting multi-token regular expressions, and allows directly specifying entity attributes, such as temporality or standard vocabulary codes. Enabling additional regular expression matching ensures greater compatibility with existing cNLP resources and knowledge bases as regular expressions are commonly used in other frameworks and languages.

UMLS Mapping

Clinical text processing systems frequently rely on comprehensive indexing of the narrative text to a standard vocabulary. The Unified Medical Language System (UMLS)\textsuperscript{27} is the most comprehensive standard terminology that is typically used as the basis for clinical concept extraction. Aiming to enable UMLS concept extraction with minimal environment configuration, medspaCy adapted QuickUMLS to spaCy framework. The existing system QuickUMLS was selected as a concept mapping solution due to its speed and matching accuracy that are comparable to other existing systems\textsuperscript{28}. It allows efficient approximate dictionary matching by leveraging an implementation of the SimString library.\textsuperscript{29} Modifications were made to QuickUMLS so that its algorithm could be contained in a spaCy component. UMLS licensing does not allow redistribution of the databases, therefore, medspaCy is distributed with resources generated from a publicly available UMLS sample.\textsuperscript{5} MedspaCy documentation includes instructions on how to create QuickUMLS resources after installation. An additional contribution of medspaCy is a better support for QuickUMLS and its dependencies for Windows operating system. While our implementation of QuickUMLS in Windows requires Anaconda and some additional installation steps, other operating systems such as Linux and MacOS can perform extractions with the provided small UMLS sample immediately following installation.

\textsuperscript{1}https://www.nlm.nih.gov/research/umls/new_users/online_learning/Meta_006.html
Contextual Analysis

Extracting contextual properties of clinical concepts is an essential step of typical cNLP systems. Negation, temporality, certainty, and experiencer are the most useful contextual features for entities extracted from clinical text. The ConText algorithm\textsuperscript{30} asserts attributes by linking entities with linguistic modifiers within a contextual window, typically a sentence. Previously implemented using Python in pyConText\textsuperscript{NLP}\textsuperscript{31}, the algorithm was adapted to spaCy\textsuperscript{API} with extended attributes. MedspaCy implements the ConText algorithm with a set of default rules that can be customized as needed. ConText rules take advantage of pattern matching and also allow the user to control the behavior of the modifier by defining properties such as directionality in the text and custom linking logic. Several attributes, such as negation, are registered by default, but attributes can be customized by users to allow for use case-specific semantic categories.

Utilities

Assisting with development of complete end-to-end systems, medspaCy provides a variety of utilities to support deploying pipelines and analyzing output of text processing:

- pre- and post-processing utilities that help developers transform simple methods into components compatible with spaCy pipelines.
- a component that converts a spaCy Doc to a dictionary-based format that is directly convertible to Pandas, SQL, CSV, or other tabular format when added to the end of an existing medspaCy pipeline. It supports output for all attributes and spans for entities, sections, and context.
- a pipeline controller for deploying medspaCy on large quantities of data. It facilitates batching documents on both input and output and a wrapper that handles standard database operations through \texttt{pyodbc} (Python ODBC bridge), such as creating tables and write queries given a set of desired columns, which allows medspaCy to deploy faster.

Visualization

The visualization tools included in spaCy’s have been adapted to display the results of the integrated medspaCy components. There are two types of visualization included (Figure 3): the first highlights extracted entities, contextual modifiers, and section titles in the text. The second represents the ConText algorithm by drawing directed arrows between entities and their linked modifiers. These visualizations can be used to present a model’s final output or to explain the logic of a model. During development, visualizing intermediate results can assist with debugging. Updating a model, testing on text, and inspecting the results with the visualizer can be done rapidly when used in an interactive setting such as Jupyter Notebook.

Use Cases

Multiple operational and research oriented projects within the U.S. Department of Veterans Affairs (VA) have utilized medspaCy. The rapid development supported by medspaCy has been essential to several projects aiding the VA’s COVID-19 pandemic response.
**Chief Complaint Surveillance**

One of the earliest applications utilizing some of these components in a spaCy pipeline is a system which is used for enhanced syndromic surveillance in VA. This system processes presenting symptoms recorded during emergency department triage such as "n/v/d" (i.e. “nausea / vomiting / diarrhea”) or "c/o cp+sob" (i.e. “complains of chest pain and shortness of breath”) to extract UMLS concepts. This pipeline has been running in an operational capacity for biosurveillance since early 2019. It was leveraged in early 2020 in response to the COVID-19 pandemic to identify potential patients under investigation before the existence of laboratory testing for SARS-CoV-2 or the term COVID-19 was adopted. This pipeline performs text pre-processing using a subset of regular expression patterns from the Emergency Medical Text Preprocessor (EMT-P)\(^{32}\), and maps text spans to standard vocabulary concepts using the QuickUMLS component.

**COVID-19 Surveillance**

Another operational project used medspaCy to identify VA patients who have been tested for COVID-19 outside of the VA network when no structured lab results were available\(^{33}\). This pipeline classifies clinical documents which likely contain a mention of positive COVID testing so that they can be prioritized for manual chart review. Between January 1 and June 15, 2020 the system had processed documents for over 3 million patients and had resulted in over 36% of all identified COVID positive patients being identified by this method alone. An evaluation of the system demonstrated precision and recall of 82.5% and 94.2%, respectively. As terminology related to the virus changed over time, it was imperative to add to concepts and patterns nearly every day. Iterative development of this system with medspaCy permitted putting the system into practice by January 21, 2020 and has been continuously operational throughout the pandemic, having processed over 63 million clinical documents one year after initial deployment.\(^8\)

**VA COVID-19 Screening**

As the COVID-19 pandemic spread, VA facilities implemented standardized COVID-19 screenings for all visitors and employees. All interactions with Veterans are logged using a specific document format that quickly became one of the most frequent documents entered into the VA EMR in 2020. Veterans in inpatient care or community living centers often had multiple COVID-19 screenings recorded each day.

A medspaCy pipeline was developed to identify the screening as a standalone document or as a subsection of a larger document and then to extract up to 17 different possible screening questions and the associated responses including community- or travel-based exposure to COVID-19, previous COVID-19 tests or diagnoses, or specific COVID-19 symptoms.

This pipeline has processed 6.8 million screening documents for 914,000 Veterans who have received a COVID-19 test. The dataset is available to VA researchers in the VA COVID-19 Shared Data Resource.\(^{34}\)

**Inpatient Nursing Templated Text Processing**

Several national VA standard operating procedures have been put into operation during the COVID-19 pandemic mandating use of standardized inpatient nursing shift assessments. While the assessments are presented to users as forms, the data are stored as the raw text conversions of templated questionnaires with their answers, thus, much of the information is not automatically retrievable. The medspaCy sectionizer is being utilized for the automatic extraction of information from these semi-structured documents (Figure 4). Given an empty version of each template, a utility module is used to automatically generate sectionizer rules. The sectionizer applies those rules to identify each question in the template as section header and extracts of all template answers captured as the section body.

**Veteran Suicide Risk and Intervention Analyses**

Several capabilities of medspaCy were used to rapidly summarize insights from Veterans, family members and organizations who responded to a request for information on how to improve suicide prevention. This survey was performed as part of the President’s Roadmap to Empower Veterans and End a National Tragedy of Suicide (PREVENTS) task force to better understand risk factors and approaches to end Veteran suicide\(^{35}\). A total of 9,040 open ended responses

\(^8\)https://spacy.io/universe/project/cov-bsv
needed to be analyzed and their insights summarized to allow a report of findings to be delivered to the task force within 3 months. Several capabilities in medspaCy made this project possible within the requested deadline, such as iterative rule development, exploratory data analysis, and visualization of text extraction. The project was also able to leverage existing knowledge resources from UMLS, including concepts related to mental health, medication and treatment. These insights were utilized by the task force as material to inform development of the PREVENTS.

**Homelessness and Housing Stability Identification**

Social determinants of health including homelessness are important factors in patient health and are often only documented in free text. The VA partners with community organizations through the Supportive Services for Veteran Families (SSVF) program to provide rapid rehousing and temporary financial support to veterans who are homeless or at risk of becoming homeless. A pipeline is being developed using medspaCy which extracts mentions of homelessness and housing stability from clinical texts. This information is then aggregated to infer a patient’s housing stability over a period of time and will be used to evaluate long-term outcomes of SSVF participants and identify at-risk individuals who could benefit from enrollment in the program.

**Education**

Besides supporting operational or clinical application, medspaCy’s simple interface and low barrier to entry makes it useful for education. Local, national, and international workshops and tutorials have utilized medspaCy to illustrate clinical text processing. A recent multi-day intensive course on clinical data science targeted to medical students used medspaCy as part of its curriculum. This course introduces medspaCy in one of the first sessions so that the remainder of the workshop can be devoted to experiential learning where students create rules and modify components to develop systems for processing clinical documentation. A simple installation process and a common programming interface enables students with little programming experience to quickly apply NLP to clinical data.

**Future work**

There are many additional components and applications possible in the next steps for medspaCy. One potential area of development is releasing medspaCy pipelines which include components, resources, and pre-trained models for specific cNLP tasks. These could also be created from publicly available resources such as MIMIC-III, shared tasks such as i2b2, or ontologies constructed during previous research. Released models could include a broad range of clinical domains, including adverse drug event detection, infectious disease surveillance, and radiology reporting.

While most of the existing medspaCy components support rule-based systems, future work could include additional utilities for machine learning. This could include word embeddings loaded as part of a default pipeline or utilities for feature extraction from text. Such improvements would allow users to more fully leverage both statistical and rule-based cNLP.

**Conclusion**

By improving the accessibility to high-performance NLP solutions optimized for clinical language, we aim to improve the productivity of cNLP development and reduce development burden for projects that require the use of rule-based

---

1. https://github.com/Melbourne-BMDS/mimic34md2020_materials
NLP or the integration of rules with other NLP methods. Unlocking new data elements through clinical information extraction will enable researchers to find answers to questions previously inaccessible for investigation.

Acknowledgements
This work was supported using resources and facilities of the Department of Veterans Affairs (VA) Informatics and Computing Infrastructure (VINCI), VA HSR RES 13-457, the Veterans Health Administration (VHA) Office of Analytics and Performance Integration (API), BASIC (Biosurveillance, Antimicrobial Stewardship, and Infection Control) and by the VA HSR&D Informatics, Decision-Enhancement, and Analytic Sciences (IDEAS) Center of Innovation (CIN 13-414).

The views expressed in this article are those of the authors and do not necessarily reflect the position or policy of the Department of Veterans Affairs or the United States government.

References


A Machine Learning Pipeline for Accurate COVID-19 Health Outcome Prediction using Longitudinal Electronic Health Records

Alice Feng

1The Harker School, San Jose, California

Abstract

Current COVID-19 predictive models primarily focus on predicting the risk of mortality, and rely on COVID-19 specific medical data such as chest imaging after COVID-19 diagnosis. In this project, we developed an innovative supervised machine learning pipeline using longitudinal Electronic Health Records (EHR) to accurately predict COVID-19 related health outcomes including mortality, ventilation, days in hospital or ICU. In particular, we developed unique and effective data processing algorithms, including data cleaning, initial feature screening, vector representation. Then we trained models using state-of-the-art machine learning strategies combined with different parameter settings. Based on routinely collected EHR, our machine learning pipeline not only consistently outperformed those developed by other research groups using the same set of data, but also achieved similar accuracy as those trained on medical data that were only available after COVID-19 diagnosis. In addition, top risk factors for COVID-19 were identified, and are consistent with epidemiologic findings.

Introduction

COVID-19 is a serious respiratory disease caused by a new type of coronavirus (SARS-Cov-2) and has put a strain on the global healthcare systems. As of March 8, 2021, there are more than 116 million confirmed infections worldwide, where more than 28 million cases are in the United States with a death rate of 1.78% which translates to 505K death due to COVID-191. In addition, there are 18 million veterans in the US who are at higher risk than the general population for severe COVID-19 illness. The death rate of veteran COVID-19 patients is 4.6%2, which is significantly higher than that of the general population. Epidemiology studies have shown that there are several factors that increase a COVID-19 patient’s risk of developing into a severe disease even death3,4,5. Hence, it is of urgent need to efficiently identify and protect at-risk individuals, especially veterans, through accurate predictive models, and effectively allocate limited healthcare resource such as vaccination to mitigate the burden on healthcare systems.

Many studies on COVID-19 prediction models were published in the past year, however, most of them were prognostic models built on suspected or confirmed COVID-19 cases, and relied on COVID-19 specific medical data such as lab tests, symptoms, and chest imaging. In addition, most models only predict COVID-19 mortality risk. There have been few studies that predict several COVID-19 health outcomes using past Electronic Health Records (EHR) routinely collected from the general population6. There was no study specifically addressing the veteran population. We filled in the gaps in this study.

In this project, we designed a novel machine learning pipeline to predict COVID-19 related health outcomes and identify the risk factors of different health outcomes using the synthetic veteran EHR data provided by the VHA (Veterans Health Administration) Innovation Ecosystem and precisionFDA COVID-19 Risk Factor Modeling Challenge7. Synthetic data mirroring the real veteran population were provided for privacy and liability concerns8. The longitudinal EHR data from 147451 patients who may or may not have COVID-19 include what are routinely collected during clinic visits, such as demographic information (age, gender, race, income, address, marital status etc.), allergies, education, immunization, medication, medical histories, insurance care plans, and many other information. The predicted health outcomes included COVID-19 infection status, mortality status, controlled ventilation status, days hospitalized and days in ICU.

Methods

1. Project goal

The 147451 patients’ EHR were split into two sets with 80% used as the training data set and 20% used as the test data set. The COVID-19 related health information (COVID-19 related symptoms, diagnosis, outcomes etc.) was withheld in the test data set. Our goal was to build predictive models from the training set to accurately predict the COVID-19 health outcomes for the test set. During this process, we also identified risk factors for COVID-19 illness.
as part of feature selection for the model training where we chose the top features that contributed most to the health outcomes. In particular, we developed a unique machine learning pipeline which included data cleaning, initial feature screening and vector representation, followed by various modules for feature selection, model training, and model evaluation.

2. Data analysis and data cleaning

First, we analyzed the EHR data extensively and designed an effective data cleaning algorithm to filter out data that may result in skewed or biased model. In machine learning, it is important to do data cleaning to remove data that are not relevant or may cause bias or skew during training. Based on our data analysis, we applied a data specific cleaning algorithm to remove the following problematic training data. 85944 records remained after data cleaning.

1. 25% of patients were dead before Jan 1, 2020. We consider these patients died of reasons other than COVID-19 since the first COVID-19 case in the United States occurred after Jan 1, 2020. They were removed as they did not offer any COVID-19 related information.

2. 0.5% of the records have conflicting data types and were removed.

3. Some patients had positive then negative PCR COVID-19 test records. They were recovered COVID-19 patients. We removed ‘COVID-19 negative’ test results from these patients as recovery was not studied in this project and these ‘negative’ test results were misleading.

4. A few patients were hospitalized due to COVID-19 but lacked positive COVID-19 PCR test records. Their COVID-19 status might have been confirmed through other methods such as imaging. We consider them having positive COVID-19 tests.

5. 1% data had conflicting condition code and descriptions, and were removed.

3. Initial feature screening

Next, we did an initial feature screening on the training data. Each patient was represented by a feature vector which included patient unique ID and the features a patient had, i.e., patient information from EHR including demographics, medical conditions, etc. In order to improve model training stability and reduce computational complexity, we first selected a subset of the features based on the following rules.

1. We selected features that were most relevant. Hence, health insurance provider, revenue, payer’s information, organization were not selected.

2. We selected features that can be more readily quantified in the feature matrix. Hence medical imaging, addresses, insurance plan were not selected.

3. We removed features that have constant values across all samples.

4. We also removed features that are highly correlated to other features to reduce data redundancy.

4. Vector representation

We then transformed the raw data into feature vectors. The goal is to minimize the distance between the targeted vector and the predicted vector. Many machine learning algorithms require that the numerical values of each feature can be compared against each other and a valid distance can be defined to reflect the closeness between the values. Thus, an optimized feature vector representation is critical to achieve the optimal model performance. Based on this principle, our vector representation algorithm

1. used frequency for medical condition, allergy, medication, immunization, medical procedure;

2. used One-Hot Encoding for categorical variables such as marital status, race and gender, so as to expand those categorical values into a series of binary numbers;

3. used the latest record for ‘observations’;

4. transformed the raw birthdate into a numerical feature for age.

5. Feature normalization

Features with very large values may dominate the model training. In addition, gradient descent, used in many machine learning algorithms, converges much faster with normalized features. Hence we normalized the features by removing each feature’s mean and scaling each feature to unit variance.
6. Feature selection through ANOVA
Coupled with model training, the significance value of each feature was calculated using t-test through Analysis of Variance (ANOVA) for each health outcome and reflected the feature’s impact on the health outcome. We tested both FWE (family-wise error rate) and FPR (false positive rate) modes of ANOVA and chose the one with better performance. In the end, about half of the features were removed due to their low significance.

7. Model training
We adopted different machine learning strategies for different health outcomes.

(1) Binary status predictions such as COVID-19 infection status, mortality status and ventilation status were classification problems, and we experimented with algorithms such as Gradient Boosting Machine (GBM) classifier, AdaBoost classifier, Random Forest (RF) classifier, K-Nearest Neighbor (KNN) classifier and ensemble learning.

(2) Numerical predictions such as days hospitalized and days in ICU were regression problems, and we experimented with algorithms such as GBM regressor, AdaBoost regressor, RF regressor, KNN regressor, and ensemble learning.

We chose these machine learning strategies based on machine learning literatures where these algorithms performed well in similar settings. In general, boosting algorithms combine the results from a series of weak prediction models to make the final prediction. Random Forest combines the results of many individual decision trees to make the final decision. KNN classifies data points based on the similarity of the points.

Each algorithm has a set of hyper-parameters, which were optimized through 5-Fold Cross Validation with Grid Search. In particular, the performance of the algorithm given a grid point was evaluated through 5-Fold Cross Validation to reduce the risk of overfitting. Here, the 5-Fold Cross Validation split the training data set into 5 folds and used four folds to train the model and the remaining one fold to test the model, then rotated through the training/testing folds in five iterations. The average performance over five iterations was used as the performance of the algorithm for that grid point. After traversing all grid points, the optimal set of hyper-parameters was selected for that algorithm.

As the last step of model training, the algorithm was trained again over the entire training data using the optimized set of hyper-parameters to get the final model.

8. Model testing
Eventually, we chose three models with best cross validation performance on the training data and predicted the COVID-19 health outcomes for the testing data. For each patient in the testing data, we predicted the probability of COVID-19 infection, probability of mortality, probability of using controlled ventilation, the length of hospitalization and the length of ICU stay, due to COVID-19. The predictions were evaluated against the ground truth of the testing data. AUROC (Area Under Receiver Operating Characteristics Curve) for classifiers, RMSE (Root Mean Square Error) and C-index for regressors were used as the performance metric. ROC (Receiver Operating Characteristics) curve is a graph showing the performance of a classification model at all classification thresholds, while AUROC measures the accuracy of the classifier. C-index measures the goodness of fit for binary outcomes in a regression model. The features with top significance values were selected as the high-risk factors for each health outcome.

Results

1. Raw data analysis
Our statistical analysis of the veteran EHR training data showed the following statistics (Table 1).

<table>
<thead>
<tr>
<th>Table 1 Training data statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>number of patients</td>
</tr>
<tr>
<td>number of patients who took COVID-19 test</td>
</tr>
<tr>
<td>number of COVID-19 tests total</td>
</tr>
<tr>
<td>number of COVID-19 positive patients</td>
</tr>
<tr>
<td>number of COVID-19 negative patients</td>
</tr>
<tr>
<td>Feature category</td>
</tr>
<tr>
<td>----------------------------------------------</td>
</tr>
<tr>
<td>Medical conditions</td>
</tr>
<tr>
<td>Observations</td>
</tr>
<tr>
<td>Allergies</td>
</tr>
<tr>
<td>Medical encounters</td>
</tr>
<tr>
<td>Medications</td>
</tr>
<tr>
<td>Medical procedures</td>
</tr>
<tr>
<td>Age</td>
</tr>
<tr>
<td>Marital status</td>
</tr>
<tr>
<td>Race</td>
</tr>
<tr>
<td>Gender</td>
</tr>
</tbody>
</table>

2. Initial feature screening

After initial feature screening, 687 features (Table 2) were used in the feature vector for each patient.

Table 2 A total of 687 initial features were used in the feature matrix.

3. Model performance

For COVID-19 positive status, mortality status, and ventilation status predictions, Figs. 1-3 and Table 3 compare the ROC curves and AUROC of our models using several machine learning algorithms based on cross validation. The algorithms with the best performance were GBM, AdaBoost and an ensemble model of GBM and AdaBoost.

For days hospitalized and days in ICU predictions, the RMSE of our models are compared in Table 3. GBM, AdaBoost and an ensemble model of GBM and RF performed the best.

In Table 4, our top 3 models (marked in blue) were compared with other 31 models submitted by other research groups in terms of AUROC, RMSE and C-index. Only the top models were included for brevity.

In Tables 3 and 4, Ensemble (G+A) is Ensemble of Gradient Boosting and AdaBoost. Ensemble (G+R+A) is Ensemble of Gradient Boosting, Random Forest and AdaBoost. Ensemble (G+R) is Ensemble of Gradient Boost and Random Forest.
Table 3 Performance comparison during training for our models for each COVID-19 health outcome

<table>
<thead>
<tr>
<th>COVID-19 positive</th>
<th>Mortality prediction</th>
<th>Controlled ventilation status prediction</th>
<th>Days hospitalized prediction</th>
<th>Days in ICU prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUR OC</td>
<td>Model</td>
<td>AUROC Model</td>
<td>AUROC Model</td>
<td>RMSE Model</td>
</tr>
<tr>
<td>0.542</td>
<td>AdaBoost</td>
<td>0.842</td>
<td>Ensemble (G+A)</td>
<td>0.802</td>
</tr>
<tr>
<td>0.539</td>
<td>Ensemble (G+A)</td>
<td>0.838</td>
<td>AdaBoost</td>
<td>0.802</td>
</tr>
<tr>
<td>0.529</td>
<td>Gradient Boosting</td>
<td>0.837</td>
<td>Gradient Boosting</td>
<td>0.793</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Table 4 Performance comparison during testing for our models (marked in blue) and other groups’ models for each COVID-19 health outcome

<table>
<thead>
<tr>
<th>Mortality status prediction</th>
<th>Controlled ventilation status prediction</th>
<th>Days hospitalized prediction</th>
<th>Days in ICU prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUROC Model</td>
<td>AUROC Model</td>
<td>RMSE Model</td>
<td>RMSE Model</td>
</tr>
<tr>
<td>0.8373 Ensemble (G+A)</td>
<td>0.7976 Gradient Boosting</td>
<td>5.7637 Ensemble (G+R)</td>
<td>0.6343 Ensemble (G+R)</td>
</tr>
<tr>
<td>0.8362 Gradient Boosting</td>
<td>0.7966 Ensemble (G+A)</td>
<td>5.7643 Gradient Boosting</td>
<td>0.6332 AdaBoost</td>
</tr>
<tr>
<td>0.8361 Model 9</td>
<td>0.7939 AdaBoost</td>
<td>5.7708 AdaBoost</td>
<td>0.6323 Gradient Boosting</td>
</tr>
<tr>
<td>0.8325 AdaBoost</td>
<td>0.7923 Model 5</td>
<td>5.8176 Model 7</td>
<td>0.6122 Team 4</td>
</tr>
<tr>
<td>0.8319 Model 4</td>
<td>0.7781 Model 5</td>
<td>5.9169 Model 8</td>
<td>0.5813 Team 6</td>
</tr>
<tr>
<td>0.8117 Model 10</td>
<td>0.7762 Model 4</td>
<td>6.0081 Model 4</td>
<td>0.6004 Team 9</td>
</tr>
<tr>
<td>0.8112 Model 5</td>
<td>0.7762 Model 6</td>
<td>6.1236 Model 6</td>
<td>0.5997 Team 11</td>
</tr>
<tr>
<td>0.8094 Model 6</td>
<td>0.7671 Model 19</td>
<td>6.2020 Model 14</td>
<td>0.5908 Team 8</td>
</tr>
<tr>
<td>0.7922 Model 11</td>
<td>0.7543 Model 28</td>
<td>6.2236 Model 11</td>
<td>0.5874 Team 7</td>
</tr>
<tr>
<td>0.7888 Model 28</td>
<td>0.7370 Model 11</td>
<td>6.3318 Model 12</td>
<td>0.5830 Team 10</td>
</tr>
</tbody>
</table>

4. High-risk factors

After model training, we studied the features with high significance values. Combining medical descriptions, medications, labs, diagnosis etc., we compiled a list of medical conditions that are considered top risk factors for each health outcome in Table 5.

### Table 5 Top significance features for each COVID-19 health outcome in our best performing model

<table>
<thead>
<tr>
<th>COVID-19 positive (description, medications)</th>
<th>Mortality</th>
<th>Controlled Ventilation (description, medications)</th>
<th>Days in Hospitalization (description, medications)</th>
<th>Days in ICU (description, medications)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypertension (medication, lab)</td>
<td>Age</td>
<td>Age</td>
<td>Kidney disease (Pyelonephritis, Renal dialysis for Kidney failure)</td>
<td>Pulmonary disease (lung surgery, lung cancer)</td>
</tr>
<tr>
<td>Metabolic syndrome (high blood pressure, high blood sugar, excess body fat, abnormal cholesterol, Hypertriglyceridemia, Triglycerides)</td>
<td>Prostate cancer (medication, lab)</td>
<td>Prostate cancer (medication, lab)</td>
<td>Kidney disease (Anemia due to kidney failure)</td>
<td>Kidney disease (Anemia due to kidney failure, Renal dialysis)</td>
</tr>
<tr>
<td>Diabetes (Diabetic renal disease, Hyperglycemia, Neuropathy due to diabetes)</td>
<td>Cancer (description, medications)</td>
<td>Metabolic syndrome (abnormal cholesterol level, medications)</td>
<td>Liver disease (Total Bilirubin, AST, ALT elevated)</td>
<td>Heart disease (Catheter ablation, heart attack, Electrical cardioversion, medications)</td>
</tr>
<tr>
<td>Cancer (description, medications)</td>
<td>Metabolic syndrome (high blood pressure, abnormal cholesterol level, Triglycerides Hypertriglyceridemia, medications)</td>
<td>Diabetes (insulin injection, medications, Diabetic renal disease)</td>
<td>Gallbladder infection (Cholecystitis, Gallstones)</td>
<td>HIV positive</td>
</tr>
<tr>
<td>Chronic kidney disease (lab, medications)</td>
<td>Diabetes (insulin injection, Diabetic renal disease)</td>
<td>Hypertension (medication)</td>
<td>Diabetes (Insulin injection, Diabetic renal disease)</td>
<td>Cancer, Primary tumor</td>
</tr>
<tr>
<td>Heart disease (lab, medications, Chronic congestive heart failure, Transthoracic)</td>
<td>Coronary Heart Disease, Stroke</td>
<td>Heart disease (Triglycerides, Coronary heart disease, stroke)</td>
<td>Coronary heart disease</td>
<td>Cancer, Primary tumor</td>
</tr>
</tbody>
</table>
Table 5 continued

<table>
<thead>
<tr>
<th>COVID-19 positive</th>
<th>Mortality</th>
<th>Controlled Ventilation</th>
<th>Days in Hospitalization</th>
<th>Days in ICU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obesity</td>
<td>Alzheimer</td>
<td>Kidney disease (medications, lab)</td>
<td>Cancer (description, medications)</td>
<td>Diabetes (blindness due to diabetes, insulin injection, medications)</td>
</tr>
<tr>
<td>Alzheimer</td>
<td>Chronic kidney disease</td>
<td>Bone disease (Osteoporosis, medications)</td>
<td>Lung cancer, Pulmonary disease</td>
<td>High blood pressure (medications)</td>
</tr>
<tr>
<td>Pulmonary disease (lab, lung cancer)</td>
<td>Lung cancer</td>
<td>Alzheimer</td>
<td>Smoking</td>
<td>Bone disease (Osteoporosis)</td>
</tr>
<tr>
<td>Alzheimer</td>
<td>Anemia</td>
<td>Anemia</td>
<td>Metabolic syndrome</td>
<td>High cholesterol (medication, lab)</td>
</tr>
<tr>
<td>Tobacco smoking</td>
<td>Tobacco smoking</td>
<td>Lung cancer</td>
<td>Alzheimer</td>
<td>Alzheimer</td>
</tr>
</tbody>
</table>

Discussion

1. High-risk factors

Even though the top risk factors for each health outcome are not exactly the same, they cover the same range of issues: age, diabetes, metabolic syndromes (hypertension, high blood sugar, abnormal cholesterol levels), cancer (especially breast cancer, prostate cancer, lung cancer), heart disease, and chronic kidney disease. These risk factors are consistent with published epidemiological findings. In this study, gender, race or income did not show high significance. The reason that age did not show up as a high-risk factor for days in hospital or in ICU is because the other illnesses shown in Table 5 have a more direct impact. Being old but healthy will not cause longer hospital or ICU stays.

2. Model analysis

A total of 34 models were submitted to the VHA Innovation Ecosystem and precisionFDA COVID-19 Risk Factor Modeling Challenge. Table 4 shows that our models (marked in blue) consistently outperformed almost all other 31 models in every category (only the top models were included for brevity). Detailed results are published on the modeling challenge website. We believe the unique machine learning pipeline we developed were crucial in achieving the best models, which included thorough data cleaning, systematic initial feature screening, optimal vector representation, feature normalization, methodical choice of k-Fold Cross Validation, feature selection mode selection, and exhaustive testing of combinations of different settings.

3. COVID-19 positive status prediction

The health outcome “COVID-19 positive status” was removed from the modeling challenge, since all submitted models had low AUROC around 0.52. For example, our models could only achieve AUROC of about 0.54 (Fig 3, Table 3). This suggests that the given EHR data only is not a good indictor of a person’s chance of contracting COVID-19. Combining other information may provide better correlation such as a person’s daily activity, if wearing a mask and keeping social distance.

4. Scope of prediction

Most prior work only identified COVID-19 mortality risk factors and predicted mortality probability. However, with good accuracy, our models predicted a wider scope of health outcomes including mortality, controlled ventilation, days in hospitalization, and days in ICU. These health outcomes are of great interests to healthcare systems, for example, Veteran Health Administration, so that they can predict if their facilities have enough hospital beds, ICU beds, or ventilation equipment, and better prepare for possible shortage. The ability to predict these health outcomes is also more useful than only identifying risk factors, as these predictions are based on a patient’s holistic medical
history, demographic info and social variables; hence are more accurate and provide more information than using risk factors only.

5. **Mortality prediction**

For mortality prediction, other published COVID-19 predictive models have AUROC range from 0.69-0.9. All these models relied on symptoms, lab results, chest images after patients were diagnosed or suspected of COVID-19. Our model provides relatively high AUROC of around 0.84, was built entirely on longitudinal EHR data, without the need for confirmed or suspected COVID-19 test, and without using imaging or any COVID-19 specific tests. This makes our model more advantageous as it predicts the outcome well before a patient contracts COVID-19 and can provide protection for high-risk patients well in advance.

6. **Significance of longitudinal EHR**

As stated above, the ability to use longitudinal EHR data in accurately predicting several critical COVID-19 health outcomes make our models stand out among thousands of COVID-19 prediction studies.

7. **Reliability of our models**

The health outcome of the EHR testing data were not given to us. The AUROC, RMSE and C-index in Table 4 were calculated by VHA and precisionFDA using our predictions and the ground truth of the testing data. The good AUROC and RMSE/C-index values demonstrated the reliability of our models.

8. **Implication for veteran and general population**

The models were built on synthetic veteran EHR data generated by Synthea specifically for VHA for this challenge, hence the synthetic EHR mirrors the real veteran population in terms of demographics, medical history, social determinants, etc. Thus, our models are one of the first COVID-19 predictive models addressing the veteran population. In addition, even though the final models are data specific, our machine learning pipeline can be applied to the real patient data for the general population. In fact, our models are being validated using real patient data from VHA and precisionFDA and the results will be available soon.

9. **Future work**

We would like to work on the following topics in the future.

1. Applying our models to real veteran EHR data is underway.
2. We would like to test more machine learning algorithms to see if better performance is achievable, such as Bayesian Network and Support Vector Machines.
3. It would be interesting to divide the training data into separate age groups, for example, age 70+, 40-70, or 40-, then train the model for each age group, and see if the prediction is more accurate.
4. We are also interested in adding each person’s genomic data in the predictive model.
5. Making a software tool to calculate a person’s COVID-19 risk score would be quite useful.
7. Predict a person’s chance of catching COVID-19 using EHR and other relevant information.

**Conclusion**

In this paper, we described innovative machine learning strategies to accurately predict a holistic set of COVID-19 health outcomes and identify COVID-19 risk factors. In particular, we designed a machine learning pipeline which was trained and evaluated on synthetic EHR data. The health outcomes that can be predicted from our pipeline include mortality status, controlled ventilation status, length of hospitalization and length of ICU stay. For mortality and controlled ventilation status, our predictive models can achieve AUROC of around 0.8-0.84. For days hospitalized and days in ICU, our models can achieve RMSE of 5.7 and 1.5, C-index of 0.63 and 0.75, respectively. Our models consistently outperformed those from other research groups. The top risk factors we identified include age, diabetes, metabolic syndromes, cancer, heart disease, chronic kidney disease and smoking. Our results are helpful for providing advanced guidance to protect high-risk individuals especially veterans, and mitigate the burden on healthcare systems. This project also demonstrates that longitudinal EHR data can be successfully employed to provide a holistic prediction of an individual’s health risk based on past health records, which is critical for emerging infectious diseases control such as COVID-19.
Acknowledgement
This project was done as part of an internship at Sentieon Inc.

References
Chart Completion Time of Attending Physicians While Using Medical Scribes

Sarah T. Florig, MS¹, Sky Corby, MS¹, Nicholas T. Rosson, BS¹, Tanuj Devara, MS¹, Nicole G. Weiskopf, PhD¹, Jeffrey A. Gold, MD¹, Vishnu Mohan, MD, MBI¹, ¹Oregon Health & Science University, Portland, OR, USA

Abstract

Medical scribes have become a widely used strategy to optimize how providers document in the electronic health record. To date, literature regarding the impact of scribes on time to complete documentation is limited. We conducted a retrospective, descriptive study of chart completion time among providers using scribes at our organization. A total of 148,410 scribed encounters, across 55 different clinics, were analyzed to determine variations in chart completion time. There was a significant variance in completion time between specialty groups and clinics within each specialty. Additionally, chart completion time was highly variable between providers working in the same clinic. These patterns were observed across all specialties included in our analysis. Our results suggest a higher level of variability with respect to chart completion when utilizing scribes than previously anticipated.

Introduction

The implementation of electronic health records (EHRs) has led to many positive outcomes, including improved efficiency in healthcare delivery and enhancements to care because of provider access to complete and up-to-date information.¹ ² However, EHRs have also led to some unintended consequences like provider burnout, over-documentation, and stress.³ ⁴ Health care professionals have some of the highest levels of burnout compared to other professions and the use of EHRs has been implicated as one of the possible factors contributing to burnout.³ ⁴ ¹¹ Because of the EHR and documentation requirements, many providers experience EHR fatigue and note bloating.¹² ¹³ Providers are spending an exorbitant amount of time doing computerized provider order entry and after-hours charting; this phenomena is known as “pajama time”, where providers are spending time working, at the EHR, after business hours.³ ⁶ ⁷ ⁹ ¹⁰ One study noted that for a patient encounter, providers spend an average of 16 minutes and 14 seconds using the EHR; the average patient visit is 15 minutes, so providers are often spending more time documenting in the EHR than they are spending with the actual patient.¹⁰ The majority of this time spent in the EHR is distributed, relatively equally, around three domains: orders, documentation, and chart review. With the excessive time spent in the EHR, both during and outside of work hours, providers have issues creating a work-life balance – many wish they could spend more time with patients.³ ¹³ ¹⁴

There are multiple ways to tackle the inefficiencies of the EHR and reduce the volume of time providers spend with the EHR. Medical scribes are one proposed way to alleviate physician documentation burden.⁴ Scribes have been aiding in documentation for centuries, but were not part of the medical field until 1975 when they were deployed as nursing scribes.¹⁸ With the massive increase in EHR use in the early 2000s, the scribing industry has risen in popularity. In today’s terms, a medical scribe is usually an unlicensed member of the clinical care delivery team who documents the patient encounter for providers in real-time, so that providers can spend more time with the patient and less time in the EHR. Besides documentation, research has suggested that medical scribes can also aid in information gathering and data entry.¹⁹ It has been demonstrated that medical scribes may decrease provider EHR-time, increase patient and provider satisfaction, improve workflow efficiencies and boost billing and reimbursement through the optimization of coding.²¹ ²⁸

A systematic review and meta-analysis by Gottlieb et al. [2021] found that changes in satisfaction, patient throughput, and revenue observed with the use of scribes was nearly consistent across the literature. Their analysis consisted of 562,682 patient encounters from 39 different studies. They observed that providers could see more patients per hour and that the use of a scribe resulted in increased relative value units (RVUs), a measure of the relative economic value of the medical services provided.²⁸ Seven of the nine studies that investigated provider perceptions when working with scribes reported increases in satisfaction. Of the 18 studies that looked at the patient view of scribe use, reports had varying findings. Some reports suggested that scribes did not change patient satisfaction, while others found increases in patient satisfaction when providers used scribes. One important thing to note about this study is that even though the studies consisted of 562,682 encounters, there was such heterogeneity regarding both study design and measures used, that the authors could not perform a meta-analysis on all of the data elements, such as provider satisfaction.²⁸

While research has suggested that scribes can decrease provider burnout by decreasing EHR documentation burden, there is a gap in the literature on whether using scribes leads to better chart closure and documentation. This is of particular concern, considering that there have been mixed results as to the quality of scribed documentation. One
study showed that scribe notes have a higher documentation quality than non-scribed notes. Other studies noted variability in the accuracy of the scribed note. Some institutions have looked at chart closure time with scribe use.

One outpatient ophthalmology clinic used audit logs of EHR data to investigate the scribes’ impact on clinical documentation; their results showed that providers who used scribes had overall less documentation time, but were spending more time documenting after the visit when using a scribe. Another study conducted in an outpatient oncology department reported that providers who used scribes spent less time documenting at the end of the day compared to providers who did not have scribes. A recent meta-analysis found that scribes had a varying impact on documentation completion, with studies ranging from positive to no impact. In another study, there was no difference between providers who used scribes and providers who did not use scribes when it came to chart notes that were incomplete after 72 hours. One potential reason for these varied results is that the vast majority of the studies investigating chart closure time only include one or very few clinics or specialties. These studies may not account for varying workflows and thus researchers cannot compare or contrast chart completion time between different groups. Thus, there is a need to conduct large-scale quantitative studies using multiple clinics, specialties, and subspecialties that examine how scribes impact provider chart closure time.

Oregon Health & Science University (OHSU) piloted a “home grown” scribe program in 2011 for the Center for Women’s Health. In 2015, it became a formal internal scribe program, which has since expanded to include over 80 clinics across OHSU. The size, duration, and breadth of the program is uniquely positioned to perform a large-scale quantitative analysis of the impact of scribes on provider documentation patterns. Thus, the goal of this study was to use EHR data to determine the impact of scribes on chart closure time across the institution, and determine factors associated with differences in completion time.

Methods

Setting and Participants. This study was conducted at a large academic medical center in Portland, Oregon. We included data from the EHR (EpicCare; Epic Systems Corporation, Verona, WI) for all ambulatory encounters that occurred between 2015 through 2019, where a medical scribe contributed to documentation. The OHSU Scribe Program almost exclusively services outpatient clinics. Thus, we only included encounters that occurred in outpatient settings. We excluded encounters if they remained unclosed at time of data collection. Encounters scheduled on Saturday or Sunday were excluded from our sample to minimize the effect of potential workflow differences between weekday and end weekend clinic service. We excluded scribes coupled with advanced practice providers because of the small sample size. Finally, to minimize irregularities due to providers who do not regularly see patients, we only included encounters belonging to physician-scribe dyads if they had completed documentation for at least 100 encounters together. This study was reviewed and approved by the institutional review board at OHSU (STUDY00017599).

Data Extraction and Processing. We extracted encounter log data from the EHR, for all ambulatory documentation, where actions on patient records were linked to an EpicCare user identifier (ID) of a medical scribe. We collected the following encounter-level concepts: patient ID, visit ID, physician ID, and scribe ID. Additional meta-data about the specialty and subspecialties of outpatient clinic providers were also extracted. We created a categorical variable to identify the outpatient clinic specialty groups included in our analysis: medicine, surgery, obstetrics and gynecology (Ob-Gyn), and pediatrics. These groups were defined based on high-level differences between the workflows of the specialties. We then generated a set of identifiers to capture the nesting of scribes within physicians, the physicians within clinics, and the clinics within specialty groups.

Descriptive Statistics. For each nested level of analysis, we calculated the median and interquartile range of time to complete chart notes by finding the difference between the scheduled visit date and the date the encounter documentation was completed by the physician. Finally, we also determined the percentage of charts that were completed according to organizational policies: “on time” documentation was closed in less than 14 days, “late” documentation was closed within 14 to 28 days, and “delinquent” documentation was completed sometime after 28 days.

Statistical Analysis.

Descriptive statistics were provided for physicians, scribes, and physician-scribe dyads across our entire sample (Table 1). Continuous variables with normal distribution were presented as mean ± standard deviation, while non-normal variables were reported as median [interquartile range]. We then included data from medicine specialty clinics to illustrate the trends repeatedly observed for documentation completion time across all specialties and clinics in our sample (Figures 1 and 2). Kruskal-Wallis tests were used to compare the mean chart completion time of three or more
groups (i.e. specialty, clinic, and provider). Where significant, a Dunn’s post hoc test was carried out on each pair of specialty groups. P-values were adjusted using Bonferroni correction. The frequencies of categorical variables (chart closure type) were compared using Fisher’s exact test, when appropriate. Descriptive and inferential statistics were performed in R v4.0.3 (R Foundation for Statistical Computing, Vienna, Austria). Unless otherwise noted, we set a level of significance of p < 0.05 for all hypothesis testing.

**Results**

**Sample Characteristics.** The purpose of this study was to examine chart completion time of attending physicians using medical scribes across a variety of clinical environments. The final sample consisted of 148,410 ambulatory encounters across 55 different outpatient clinics (Table 1). Care clinics were grouped into one of four general specialties which included medicine (69,209; 47%), surgery (48,448; 32.5%), Ob-Gyn (21,326; 14.5%), and pediatrics (9,427; 6%). The types of visits accounted for in each specialty group are included in Table 2.

Our analysis included a total of 129 physicians and 127 scribes (Table 1). On average, physicians had worked with 3±2 scribes in our sample. Scribes had a mean employment length of 14±10 months and worked with anywhere from 1 to 9 physicians. We included a total of the 325 physician-scribe dyads in our final analysis. Of these dyads, 161 (50%) worked in medical specialties clinics, 76 (23%) worked in surgical, 65 (20%) worked in Ob-Gyn, and 23 (7%) worked in pediatrics. On average, dyads worked together over a 10 [6, 17] month period and documented 296 [159, 525] encounters together.

**Chart Completion Time.** Time to complete chart notes was highly right-skewed and observed a log-normal distribution. Less than half of all chart notes (72,306; 49%) were complete within 24 hours. The median number of days to complete encounter documentation was 0.95 [0.11, 5.9] days and ranged from 0.00069 to 854 days. Finally, 132,700 (89%) of the 148,410 charts were closed on time according to organizational policies. Of the remaining, 11,734 (8%) of the completed records were considered late, while 3,976 (3%) were delinquent.

**Chart Completion Time by Specialty and Clinic.** Chart completion time was most consistent (least spread) in pediatrics with a median chart completion time of 1 [0.3] days (Figure 1). Surgery had the lowest median completion time (0 [0.5] days), while Ob-Gyn had the highest (1 [0.10]). Difference in mean ranks of time to complete chart notes were highly statistically

### Table 1. Counts of the Unique Levels of Each Categorical Variable.

<table>
<thead>
<tr>
<th>Variable</th>
<th>No.¹</th>
</tr>
</thead>
<tbody>
<tr>
<td>Encounters</td>
<td>148,410</td>
</tr>
<tr>
<td>Patients</td>
<td>64,514</td>
</tr>
<tr>
<td>Physician-Scribe Dyads</td>
<td>325</td>
</tr>
<tr>
<td>Scribes</td>
<td>127</td>
</tr>
<tr>
<td>Providers</td>
<td>129</td>
</tr>
<tr>
<td>Clinics</td>
<td>55</td>
</tr>
</tbody>
</table>

¹Number of unique values.

### Table 2. Visit Types Across Clinical Specialty Groups.

<table>
<thead>
<tr>
<th>Specialty</th>
<th>Office visit</th>
<th>Procedure</th>
<th>Prenatal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medicine</td>
<td>59,499 (40)</td>
<td>7,078 (5)</td>
<td>2,632 (2)</td>
</tr>
<tr>
<td>Surgery</td>
<td>47,820 (32)</td>
<td>628 (0.5)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Pediatrics</td>
<td>9,425 (6)</td>
<td>2 (0)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Ob-Gyn</td>
<td>12,996 (9)</td>
<td>702 (0.5)</td>
<td>7,628 (5)</td>
</tr>
<tr>
<td><strong>All</strong></td>
<td><strong>129,740 (87)</strong></td>
<td><strong>8,410 (6)</strong></td>
<td><strong>10,260 (7)</strong></td>
</tr>
</tbody>
</table>

Abbreviation: Ob-Gyn, obstetrics and gynecology.

**Figure 1.** Days to Complete Chart Notes by Specialty Group.

Abbreviation: Ob-Gyn, obstetrics and gynecology.

The upper and lower ends of the boxes indicate the first and third quartiles. The horizontal line inside the box indicates the median and the square indicates the mean. The whiskers indicate values within 1.5x the interquartile range from the upper or lower quartile (or the minimum and maximum value if within 1.5x the interquartile range of the quartile).

The mean ranks of time to complete chart notes were different between groups (p < 0.0001). Brackets over bars indicate specialty groups that were statistically significant; **** indicates p < 0.0001.
significant between specialty groups (p < 0.0001) with significant follow-on differences between each of the group-pairs (p < 0.0001).

We next sought to determine whether a similar variance was present within a given department. Overall, each specialty demonstrated a high degree of variance between their subspecialty clinics. As an exemplar, data from the medical specialty clinics are found in Figure 2. Here, the number of days to complete chart notes was also highly variable. The

**Figure 2.** Days to Complete Chart Notes Within Medicine Specialty Clinics.
Each medicine clinic is represented on the x-axis by their ID (M1-25). The dashed, black line represents the overall sample mean.
Statistical components of the boxplots are explained in the first footnote to Figure 1.

The median number of days to complete chart notes ranged from 0.03 to 18 days. For some clinics, like M22 and M23, the distribution of days to complete chart notes was highly skewed, with a substantial number of encounters closed after an extended period.
We next wished to determine whether this variance existed within given outpatient clinics. To explore the differences in completion time between providers within the same clinic, we limited the analysis to specialty clinics with at least five scribe-using providers (Figure 3). Overall, differences in mean ranks of completion time for these four clinics were highly statistically significant between groups (p < 0.0001) and multiple comparison tests suggested that each clinic’s completion time was statistically significantly different from each of the others (all adjusted p-values were <

**Figure 3.** Days to Complete Chart Notes for Physicians Within Medicine Specialty Clinics of Five or more Providers.
Each medicine clinic is shaded in gray and represented above the x-axis by their IDs (M4, M10, M11, and M18 correlate to Figure 2). Each provider within a given clinic is represented on the x-axis by their ID (P1).
Statistical components of the boxplots are explained in the first footnote to Figure 1.
More importantly, within each clinic, there was a highly statistically significant difference in completion time between providers (p < 0.0001 for each clinic), suggesting that the observed variability seen across clinics is driven by variability at the individual provider-level.

Visit Day of Week & Days to Complete Chart Note. We next sought to determine what factors may explain some of the inter- and intra-provider variance and initially focused on the encounter day of the week. The total number of encounters that occurred each weekday was relatively similar. The highest number of encounters took place on Thursday (35,244; 24%), followed by Tuesday (33,100; 22%). In contrast, the lowest number of encounters occurred on Fridays (24,062; 16%), Monday (27,848; 19%), and then Wednesday (28156; 19%). The number of days to close chart notes was statistically significantly different across each day of the week that encounters were scheduled (p < 0.0001). Encounters that occurred on Thursdays and Fridays had a notably higher average number of days to complete chart notes than the other days of the week (Figure 4). Encounters that occurred on Thursdays and Fridays also had lower percentages of chart notes that were not completed “On Time” compared to the other days of the week (Table 3).

Chart Closure Day of Week & Days to Complete Chart Note. The number of charts closed each day of the week decreased throughout the workweek: Monday (29,646; 20%), Tuesday (28,969; 20%), Wednesday (27,7701; 19%), Thursday (25,4151; 17%), and Friday (19,1761; 13%). Providers can complete documentation on the weekends and our results suggest that around one in ten chart notes were completed on either Saturday (6,450; 4%) or Sunday (10,984; 7%). For charts where documentation was completed on a Saturday or Sunday, the number of days to complete the chart note was nearly double that of charts completed during the week (Figure 5). When chart documentation was completed on a Saturday or Sunday, the number of delinquent or late closures was higher than all other days of the week (Table 3).

Discussion

Previous literature regarding the use of medical scribes support their implementation as they are purported to help mitigate providers’ burdening challenges of EHR documentation.27 Despite the appeal of their intended usefulness as documentation assistants, qualities of the timeliness of documentation in the company of a medical scribe remains largely unreported. This study is among the first to quantitatively capture the variation in documentation completion time among physician-scribe dyads and the first to do so over such a wide collection of medical specialties and subspecialties. It is also the first

Table 3. Classification of Chart Closure, by Encounter Day of Week and by Chart Closure Day of Week.

<table>
<thead>
<tr>
<th>Event by day of week</th>
<th>On time</th>
<th>Late</th>
<th>Delinquent</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Encounter occurred</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Monday</td>
<td>26,079 (93)</td>
<td>1,359 (4.8)</td>
<td>718 (2.6)</td>
</tr>
<tr>
<td>Tuesday</td>
<td>31,114 (94)</td>
<td>1,660 (5.0)</td>
<td>326 (1.0)</td>
</tr>
<tr>
<td>Wednesday</td>
<td>25,518 (92)</td>
<td>1,951 (7.0)</td>
<td>379 (1.4)</td>
</tr>
<tr>
<td>Thursday</td>
<td>30,517 (87)</td>
<td>3,094 (8.8)</td>
<td>1,633 (4.6)</td>
</tr>
<tr>
<td>Friday</td>
<td>19,472 (81)</td>
<td>3,670 (15)</td>
<td>920 (3.8)</td>
</tr>
<tr>
<td><strong>Chart closure</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Monday</td>
<td>26,264 (89)</td>
<td>2,540 (8.6)</td>
<td>842 (2.8)</td>
</tr>
<tr>
<td>Tuesday</td>
<td>26,081 (90)</td>
<td>2,263 (7.8)</td>
<td>625 (2.2)</td>
</tr>
<tr>
<td>Wednesday</td>
<td>25,763 (93)</td>
<td>1,587 (5.7)</td>
<td>420 (1.5)</td>
</tr>
<tr>
<td>Thursday</td>
<td>23,435 (92)</td>
<td>1,611 (6.3)</td>
<td>369 (1.5)</td>
</tr>
<tr>
<td>Friday</td>
<td>18,041 (94)</td>
<td>865 (4.5)</td>
<td>270 (1.4)</td>
</tr>
<tr>
<td>Saturday</td>
<td>5,044 (78)</td>
<td>838 (13)</td>
<td>568 (8.8)</td>
</tr>
<tr>
<td>Sunday</td>
<td>8,072 (73)</td>
<td>2,030 (18)</td>
<td>882 (8.0)</td>
</tr>
<tr>
<td>All</td>
<td>132,700 (89)</td>
<td>11,734 (7.9)</td>
<td>3,976 (2.7)</td>
</tr>
</tbody>
</table>

Closure type was defined by institutional policies as:
On time, 0 to 14 days; Late 14 to 28 days; Delinquent +28 days.
to explore how documentation completion time manifests across multiple levels of a nested sample of clinics, physicians, and scribes, which provides insight into where causes of variation may arise.

One of the most interesting findings is the wide variance between the major clinical specialty groups and time to complete chart notes. The inconsistency in documentation completion time throughout these groups is not surprising considering the previously reported differences in workflows and EHR use among specialists and primary care physicians. Among scribe users, this variation is also understandable, given that the previous literature suggests that different specialties may not utilize their scribes in the same way. When examining documentation completion time between clinics within specialties, we found that this variability existed within the different subspecialty clinics and providers using scribes within the same clinic. This emphasizes that a large degree of variance that we observed is most certainly linked to characteristics at the individual provider level. Our findings highlight a need to develop strategies to manage variation among scribe users. Organizations need provider-specific training that can be applied during scribe implementation and additional tools, such as a dashboard or data visualization system, which can allow them to identify ineffective users of scribes for intervention.

Many factors may influence the widely observed provider differences in chart completion time. One factor that we were able to consider in this analysis was whether documentation completion time was different depending on the day of the week that a visit takes place or the day of the week on which charts were closed. Day of the week may affect completion time as some providers may be more or less likely to perform documentation activities after-hours or on weekends. Our findings suggest that charts are completed at different rates of time depending on the day of the week of a visit. Furthermore, charts that are closed later in the workweek will have an increasingly higher number of days to complete documentation than those completed earlier in the week. More importantly, our data suggests that at least 10% of scribe-generated chart notes are still being completed during “pajama time”. This significant fraction of notes, finalized on weekends, had the longest amount of time between the visit date and the closure of chart notes – documentation completed on Sunday took over double the time to complete chart notes compared to those closed during the workweek. It should be noted that this is currently only assessing weekend “pajama time” and we did not assess after-hours weekday work, which was likely also present, suggesting this may be an even greater effect.

However, the idea that, even with scribe utilization, providers are facing after-hours documentation may explain why there are conflicting results on reports of how scribe use impacts provider documentation behaviors.

It is likely that additional factors are influencing the large degree of variance with chart completion time among providers who use scribes. The degree to which providers must correct a scribe-generated note is another likely factor influencing the variance we observed in this sample. First, there may be underlying variability in the quality of the scribe-generated note. Previous simulation studies have suggested that scribe-generated notes contain a wide variance in both note structure and content, as well as discrepancies in their accuracy. While all scribe-generated documentation requires a thorough review by the supervising physician, the inconsistency of notes produced by scribes may inevitably influence the amount of time a physician must expend to review, edit, and sign-off on their chart notes. This is further complicated by previous literature regarding the underlying provider-level variance in the specific content, type of content, and amount of content that is included in encounter chart note documentation. Workflow analysis may be another useful tool that can be used in future studies to ascertain granular details of the influence that scribes have on provider chart hygiene.

Perhaps the largest driver to the degree of oversight providers maintain over scribe note content is the nature of the specific relationship they have with their scribe. Previous work carried out by our group and others have found numerous subthemes that play into this relationship and allow for effective and efficient interactions between the scribe and provider. One critical subtheme that has been identified in the research is the “quality of scribe-provider relationship” and the aspect of trust. Findings suggest that the longer the provider and scribe work together, the higher the level of trust in the relationship can become. This higher degree of trust may be associated with less oversight in the completion and correction in content of the note. Future research should investigate if and how scribe-generated chart notes differ regarding both content and quality of the documentation. Additionally, it would be helpful to determine if the content or quality of the documentation depends on the amount of time taken to complete the chart note.

It is helpful to consider these findings in the context of the limitations of this study. First, our analysis only includes a single, academic medical center, limiting the generalizability of our findings. Likely, the incentives that underlie physician documentation hygiene at an academic medical center are different from that of private practice physicians. For example, the salary of physicians working in private practice may be directly tied to reimbursement claims, which require a completed chart note before the submission of a claim is possible. For academically based physicians, salary
is often tied directly to faculty appointments as opposed to claims reimbursement, which may influence their timeliness to complete documentation. Furthermore, our institution recruits and trains their scribes, who are almost uniformly pre-professional students. Individuals from many different professional groups can serve as scribes (medical assistants, nurses, etc.), and further, a large fraction of scribes are trained and supplied by independent organizations. It will be important in future studies to determine the impact these factors have on completion time and documentation hygiene. This study did not account for the chart completion behaviors of physicians before receiving a scribe. It is possible that some of the physicians who received scribes already had poor documentation hygiene, and thus self-selecting for those providers most likely to be given a scribe in the first place, which may have influenced some of the variance observed in our findings. Further, we did not directly analyze how much of the variation we observed in this study occurred because of the presence of a scribe, and we were unable to account for the impact of confounders. Therefore, while multiple explanatory factors may influence chart completion time, there is a chance that the results presented here are attributable to other ongoing factors that were not captured by this analysis. As a result, it is difficult to know how much of a provider’s documentation completion time is associated with the presence of a scribe or these other unaccounted-for factors.

Conclusion

This study assessed a gap in knowledge on how quickly physicians complete documentation while using medical scribes. Across clinical specialties and outpatient clinics, there was substantial variation in the time to complete encounter documentation, and this variation persisted across physicians using scribes within the same clinic. Our findings suggest that individual provider behavior may drive the variation in completion of clinical documentation and that scribes may have little impact on regulating the time to chart completion. Because this variation has the potential to undermine the justifications for the use of scribes, it is important that scribe-users understand this aspect of the physician-scribe dynamic and that interventions are developed to educate scribe-users to take full advantage of the assistance with clinical documentation. It should be noted that clinical documentation in the United States is, on average, almost four times as long as those in other countries; an important factor that has driven the utilization of scribes. This work also brings to the fore a larger issue of institutional and regulatory requirements for documentation. Organizations must consider how matching physician-scribe dyads will achieve the overall goal of scribe implementation, while also recognizing the limitations that scribes will have in altering poor EHR users’ underlying behavioral deficits.

Acknowledgment and Funding

We wish to acknowledge Dr. Joan S. Ash, James Becton, Robby Bergstrom, and Dinesh Neelapala for their contributions to this research. This project is supported under contract grant #R01HS025141 from the Agency for Healthcare Research and Quality (AHRQ), U.S. Department of Health and Human Services. The results from this paper are from the authors and do not represent the views of AHRQ or the U.S. Department of Health and Human Services.
References

39. OHSU Medical Scribe Program. Our Program [Internet]. Portland, OR: Oregon Health & Science University; c2020 [cited 2021 Jan 20]. Available from: https://www.ohsu.edu/medical-scribes/our-program
Parsing Immune Correlates of Protection Against SARS-CoV-2 from Biomedical Literature

Sydney L. Foote, MPH1*, Sara Jones, MS1*, Jane Lockmuller, MS1, Liliana Brown, PhD2, Joseph Breen, PhD3, Anupama Gururaj, PhD3

1Office of Data Science and Emerging Technologies, NIAID, NIH, Rockville, MD, USA
2Division of Microbiology and Infectious Diseases, NIAID, NIH, Rockville, MD, USA
3Division of Allergy, Immunology, and Transplantation, NIAID, NIH, Rockville, MD, USA

* Both authors contributed to the work equally.

Abstract

After the emergence of severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) in 2019, identification of immune correlates of protection (CoPs) have become increasingly important to understand the immune response to SARS-CoV-2. The vast amount of preprint and published literature related to COVID-19 makes it challenging for researchers to stay up to date on research results regarding CoPs against SARS-CoV-2. To address this problem, we developed a machine learning classifier to identify papers relevant to CoPs and a customized named entity recognition (NER) model to extract terms of interest, including CoPs, vaccines, assays, and animal models. A user-friendly visualization tool was populated with the extracted and normalized NER results and associated publication information including links to full-text articles and clinical trial information where available. The goal of this pilot project is to provide a basis for developing real-time informatics platforms that can inform researchers with scientific insights from emerging research.

Introduction

Immune correlates of protection (CoPs) are measurable aspects of the immune system that indicate an individual may be protected from becoming infected1. In-depth knowledge of CoPs is essential to understanding immune responses to infectious diseases. CoPs can inform the efficacy of vaccine trials or provide an estimate of individual level immunity after previous infections1-3. With the emergence of novel pathogens such as severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2), identification of CoPs has become increasingly important.

Over the past year since the emergence of SARS-CoV-2, new literature regarding SARS-CoV-2 and the resulting illness, coronavirus disease 2019 (COVID-19), is being published at a rapid pace3. Many studies have attempted to identify aspects of the immune system that may correlate with or directly confer protection against SARS-CoV-2, and several vaccine candidates have been approved for emergency use around the world4-8. Even with this research, the CoPs against SARS-CoV-2 remain ambiguous.

The vast amount of preprint and published literature provides an opportunity to gain a better understanding of potential CoPs against SARS-CoV-2. However, identifying publications relevant to SARS-CoV-2 CoPs using basic keyword searches oftentimes bring back thousands of results, making it challenging for researchers to easily stay up to date on information regarding CoPs against SARS-CoV-2. Several systems currently exist to extract COVID-19 and SARS-CoV-2 related literature from broader sources, but none are specific to literature relevant to CoPs9-12. There are also several open source natural language processing (NLP) models trained on biomedical text and literature13-15; however, these models have neither been trained to classify CoP relevant literature nor identify CoPs. Customization of existing machine learning methods, such as document classifiers and NLP, can be employed to fill this gap.

This pilot project aims to leverage existing machine learning tools to develop a pipeline that allows for quickly informing individuals of the current literature related to potential CoPs against SARS-CoV-2. To accomplish this, we identified literature related to CoPs against SARS-CoV-2, extracted relevant entities from the manuscripts, and visualized the results in a user-friendly tool. To the best of our knowledge, this is the first study focused on extracting and classifying potential CoPs against SARS-CoV-2 from biomedical literature. This will facilitate upkeep of knowledge on the status of CoPs against SARS-CoV-2 in both preprints and publications for researchers and administrators in related fields of study.

Methods
Data Curation

To identify potential CoPs against SARS-CoV-2, we downloaded the planned outcome measurements for the 252 COVID-19 vaccine trials documented on ClinicalTrials.gov on October 5, 2020\textsuperscript{16}. 113 of the 252 vaccine trials’ planned outcome measurements contained potential CoPs. We manually extracted and curated a list of these CoPs, which was reviewed by two subject matter experts (Table 1).

Table 1. Potential immune correlates of protection (CoPs) against SARS-CoV-2 and the count of appearances in planned outcome measurements from clinical trials extracted from COVID-19 vaccine trials on ClinicalTrials.gov\textsuperscript{16}; n = 113.

<table>
<thead>
<tr>
<th>Potential CoPs</th>
<th>Count</th>
<th>Potential CoPs</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Antibody</td>
<td>40</td>
<td>IgE Antibody</td>
<td>1</td>
</tr>
<tr>
<td>B-cell response</td>
<td>3</td>
<td>IgG Antibody</td>
<td>35</td>
</tr>
<tr>
<td>Binding antibody</td>
<td>16</td>
<td>IgM Antibody</td>
<td>9</td>
</tr>
<tr>
<td>CD4\textsuperscript{+} T Lymphocytes</td>
<td>7</td>
<td>Immune Signature</td>
<td>5</td>
</tr>
<tr>
<td>CD8\textsuperscript{+} T Lymphocytes</td>
<td>10</td>
<td>Immunological Memory</td>
<td>1</td>
</tr>
<tr>
<td>Cell-Mediated Immunity</td>
<td>14</td>
<td>Natural Killer Cells</td>
<td>2</td>
</tr>
<tr>
<td>Humoral Immunity</td>
<td>3</td>
<td>Neutralizing Antibody</td>
<td>52</td>
</tr>
<tr>
<td>IgA Antibody</td>
<td>4</td>
<td>T-Cell Response</td>
<td>21</td>
</tr>
</tbody>
</table>

We then used this list and words synonymous with protection to generate keyword search queries in the National Institutes of Health (NIH) Office of Portfolio Analysis’ iSearch COVID-19 Portfolio, an expert-curated source for publications and preprints related to COVID-19 and SARS-CoV-2\textsuperscript{11}. We accessed the iSearch COVID-19 Portfolio from 10/16/2020 to 11/4/2020 to identify articles discussing CoPs and generated a COVID-19 dataset of articles relevant and non-relevant to CoPs against SARS-CoV-2 for the classification task. We tried several search queries such as the following: \((antibod* \text{ OR } T-cell~1 \text{ OR } B-cell~1 \text{ OR } "natural killer" \text{ OR } immunoglobulin \text{ OR } Ig*) \text{ AND } (*neutraliz* \text{ OR } protect*)\).

Since most of our queries yielded thousands of results, we limited our results to articles with an assigned PubMed ID (PMID) and the source listed as “Peer Reviewed (PubMed)”. For example, the query above, when filtered to only “Peer Reviewed (PubMed)”, resulted in 4,961 articles. We focused on primary literature with original results by excluding publication types such as reviews, editorials, meta-analyses, and letters. Articles were then randomly selected for manual annotation. After reviewing the first set of 100 articles, we identified only 12 articles as relevant. Due to this low number, we then considered only articles with an iSearch generated relevance score of at least 0.4 (range 0.0 – 1.0) to increase the number of relevant articles. We also filtered to articles with an abstract length of at least 20 words and less than 1,000 words to control for unlabeled or mislabeled publication types of non-primary literature. Finally, we combined each article’s title and abstract into a single document. We will refer to document in place of title and abstract going forward.

Two annotators (SLF and SJ) read and manually labeled each document in the COVID-19 dataset as either relevant or non-relevant to the topic of CoPs against SARS-CoV-2 using doccano, an open-source text annotation tool with features such as text classification and sequence labeling\textsuperscript{17}. Inter-annotator agreements (IAA) were periodically calculated, and an annotation guideline was maintained to track the annotation process.

The IAA was initially 0.53 but grew to 0.70 after discussion between the annotators and edits of the annotation guideline. Further discussion and finalization of the annotation guideline resulted in a final IAA of 1.00.

The publication dates of the articles from iSearch ranged from November 2019 to November 2020. After manually annotating COVID-19 literature, 426 documents were included in the COVID-19 dataset with 55 labeled as relevant to CoPs against SARS-CoV-2 and 371 labeled as non-relevant. Since the COVID-19 dataset was heavily skewed towards non-relevant documents, we expanded our search for literature related to CoPs predating the pandemic from the years 2000 to 2019 to increase the number of relevant documents and reach our final COVID-19 + non-COVID-19 dataset used for machine learning classification.

We used several phrases in PubMed such as correlates of protection and vaccine immunity to find relevant primary articles. We then utilized PubTator Central, a web-based system providing automatic annotations of biomedical terms,
to download the titles and abstracts from the PMIDs of our searches\textsuperscript{18}. The process of manually annotating documents described above was repeated for this set of documents.

We identified another 458 relevant documents and 402 non-relevant documents from literature related to other diseases and published between 2000 and 2019 for the non-COVID-19 dataset. The final COVID-19 + non-COVID-19 dataset used for the classification task had 1009 documents: 513 relevant and 496 non-relevant.

\textit{Machine Learning Classifier}

We developed a machine learning model to classify documents as either relevant or non-relevant to the topic of CoPs. Each document from the final COVID-19 + non-COVID-19 dataset described above was preprocessed by separating special characters from words and converting text to lowercase using code adapted from the Convolutional Neural Networks for Sentence Classification\textsuperscript{19}. We also removed English stop words from each document using scikit-learn’s TfidfVectorizer’s stop words parameter\textsuperscript{20}.

We then applied a train-test split of 80/20 to the COVID-19 + non-COVID-19 dataset. After fitting TfidfVectorizer to the training dataset, both the training and test datasets were transformed to a matrix of term frequency-inverse document frequency (TF-IDF) features\textsuperscript{20}. The following machine learning models from scikit-learn were used: LogisticRegression, RidgeClassifier, Perceptron, PassiveAggressiveClassifier, KNeighborsClassifier, LinearSVC (L1 and L2 regularization), SGDClassifier (L1, L2, and elastic net regularization), MultinomialNB, BernoulliNB, and ComplementNB. In addition, we used an ensemble method, VotingClassifier, that uses a majority rule voting of the machine learning models just described\textsuperscript{20}. For model evaluation, we computed the accuracy and the average precision, recall, and F1 scores on the test dataset. In addition, we calculated a repeated stratified k-fold cross validation (5 folds and 3 repeats) with the test dataset and used accuracy as the evaluation metric. Hyperparameter tuning was not performed on the selected model.

\textit{Named Entity Recognition}

We built a custom named entity recognition (NER) model to identify and classify specific words and phrases in documents relating to CoPs. Through this model, we aimed to identify four unique entity classes: CoPs; assays used to measure CoPs; animal models used in the study, including human patients; and vaccines specified in the documents.

To construct a training and test dataset for model building, we manually identified and classified all pertinent entities within an NER dataset, a subset of the final COVID-19 + non-COVID-19 dataset, using doccano\textsuperscript{17}. An annotation guideline for entity recognition was written during the annotation process.

Two annotators (SLF and SJ) independently annotated 20 documents, resulting in an IAA of 0.74. The annotators discussed and resolved all discrepancies, with any changes added to the annotation guideline. The remaining documents in the NER dataset were split into two groups, annotated independently by one annotator each, then reviewed by the other annotator for consistency.

To build the customized NER model, we fed the annotated NER dataset into a standard NLP pipeline based upon the “en\_core\_sci\_lg” model from the Python package scispaCy and tweaked using additional tools from the Python package spaCy\textsuperscript{15, 21}. The annotated NER dataset was pre-processed using a tokenizer, parts-of-speech (POS) tagger, and parser, then piped through the customized NER. During training and testing, precision, recall, and F1 scores were used for model evaluation.

Several steps were taken to optimize the NER model. First, we added extra suffixes ([,], [0-9], [-], [.] and infixes ([ ] to the tokenizer to fix a misalignment problem with the text. We tested two different training options for the model: k-fold cross validation and an 80/20 train-test split of the NER dataset. We also varied the number of iterations used during training to find the best model fit. Additionally, we added a post-processing rule after the NER to identify any mention of “human(s)” and classify it as an “animal” entity because “human(s)” was not originally included when annotating for the NER.

For the final processing, text was fed into the NLP pipeline. The tokenizer, POS tagger, and parser pre-processed the text, then the customized NER extracted and classified entities of interest. The entity ruler identified any instances of “human(s)” and the model outputted annotated text, with the entities of interest identified and classified (Figure 1).
Interactive Visualization

In collaboration with the Analytics, Visualizations, Insights, & Data Science team within the Office of Cyber Infrastructure and Computational Biology (OCICB) at the National Institute of Allergy and Infectious Diseases (NIAID), NIH, we built an interactive Tableau dashboard to highlight the output of our models.

To populate the Tableau dashboard, we downloaded the documents and their metadata from all preprint and published papers in the iSearch COVID-19 Portfolio published between 11/1/2019 and 12/31/2020 (accessed 12/31/2020). We ran these iSearch documents through our finalized classifier and NER models. Only documents that had a predicted probability score of at least 0.5 from the classifier were considered relevant to CoPs against SARS-CoV-2 and, thus, kept for additional analysis. We used the default predicted probability score of 0.5 as our minimum threshold since this value captured all the relevant documents in a dataset consisting of only COVID-19-related literature. To allow for easy visualization, we cleaned all entities to reflect a more harmonized term using regular expressions and fuzzy string matching. For example, all instances of mouse-related entities, such as mice, humanize mouse, etc., were categorized simply as “mouse”. To accurately clean the vaccine entity data, we used categories identified in the Milken Institute Vaccine Tracker. Regular expressions were also used to extract any clinical trial IDs in the documents. Hyperlinks were generated for DOIs, PMIDs, and clinical trials IDs.

Software

All analyses were conducted using Python version 3.7.3, scikit-learn versions 0.23.2 – 0.24.1, scispaCy version 0.3.0, spaCy version 2.3.2, and Tableau Server version 2020.2.4. Training of the classifier and NER models were done on NIH and NIAID high performance clusters (HPC), Biowulf and Locus respectively.

Results

Classifier for Information Retrieval

We used the final COVID-19 + non-COVID-19 dataset to test several classifiers and determined accuracy for the various algorithms. All the tested classifiers and their performance are enumerated in Table 2. The ensemble majority voting classifier performed similarly to other models such as the SGDClassifier with L2 (accuracy of 98.51% and 98.02%, respectively). The models used for the classification task all had a median accuracy of at least 85% (Figure 2). We had several models with high accuracy and consistent performance from repeated stratified k-fold cross validation. Therefore, we decided to use the SGDClassifier with L2 regularization for all subsequent steps.

Table 2. Model evaluation measures for classification task. Precision, recall, and F1 scores were calculated for each class to find the unweighted means.

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy (%)</th>
<th>Precision (%)</th>
<th>Recall (%)</th>
<th>F1 Score (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RidgeClassifier</td>
<td>98.02</td>
<td>98.01</td>
<td>98.01</td>
<td>98.01</td>
</tr>
<tr>
<td>Perceptron</td>
<td>96.04</td>
<td>96.02</td>
<td>96.02</td>
<td>96.02</td>
</tr>
<tr>
<td>PassiveAggressiveClassifier</td>
<td>97.52</td>
<td>97.46</td>
<td>97.62</td>
<td>97.52</td>
</tr>
<tr>
<td>KNeighborsClassifier</td>
<td>84.16</td>
<td>86.45</td>
<td>85.05</td>
<td>84.08</td>
</tr>
<tr>
<td>LinearSVC with L2</td>
<td>97.52</td>
<td>97.49</td>
<td>97.55</td>
<td>97.51</td>
</tr>
<tr>
<td>LinearSVC with L1</td>
<td>98.51</td>
<td>98.45</td>
<td>98.61</td>
<td>98.51</td>
</tr>
<tr>
<td>SGDClassifier with L2</td>
<td>98.02</td>
<td>97.97</td>
<td>98.08</td>
<td>98.01</td>
</tr>
</tbody>
</table>
SGDClassifier with L1
98.51  98.45  98.61  98.51
SGDClassifier with Elastic Net
98.51  98.45  98.61  98.51
MultinomialNB
95.05  95.09  95.30  95.05
BernoulliNB
96.04  95.98  96.09  96.03
ComplementNB
94.55  94.64  94.84  94.55
Logistic Regression
97.52  97.49  97.55  97.51
Ensemble Majority Voting
98.51  98.48  98.54  98.51

Figure 2. Accuracy of machine learning models from repeated stratified k-fold cross validation on classification task.

We assessed the SGDClassifier model on the COVID-19 dataset, with 371 non-relevant and 55 relevant documents. Note that the model was trained on all 55 relevant documents and 94 non-relevant documents from this dataset. Our model achieved an accuracy of 90.1%. The model identified all the relevant documents in the COVID-19 dataset. Of the non-relevant documents, the model incorrectly classified 42 documents as relevant (Table 3). Overall, the performance of the model on correctly identifying the relevant documents had a precision of 56.70%, recall of 100%, and F1 score of 72.37%.

<table>
<thead>
<tr>
<th>Actual Class</th>
<th>Positive</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>True Positive: 55</td>
<td>False Positive: 42</td>
</tr>
<tr>
<td>Negative</td>
<td>False Negative: 0</td>
<td>True Negative: 329</td>
</tr>
</tbody>
</table>

Table 3. Confusion matrix of SGDClassifier with L2 regularization on the COVID-19 dataset (55 relevant and 371 non-relevant).

NER
We trained and tested the final NER model using 165 relevant documents from the final COVID-19 + non-COVID-19 dataset; 105 of these documents were related to COVID-19, while the remaining 60 were related to other diseases.
From these documents, we manually annotated 184 assays, 278 animals, 515 vaccines, and 1290 correlates. To build the model, we split these documents so 80% were in the training dataset and 20% were reserved for testing. We trained the model over 100 iterations of the training dataset.

We evaluated NER model performance using F1 scores, precision, and recall for the overall model as well as each entity class (Tables 4 & 5). Model performance was most affected by number of publications (Table 4).

**Table 4.** Model evaluation measures for named entity recognition (NER).

<table>
<thead>
<tr>
<th>Approach</th>
<th>Number of Training Documents</th>
<th>Number of Iterations</th>
<th>F1 Score (%)</th>
<th>Precision (%)</th>
<th>Recall (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>80/20 Train-Test Split</td>
<td>165</td>
<td>100</td>
<td>74.01</td>
<td>76.32</td>
<td>71.85</td>
</tr>
<tr>
<td></td>
<td>55</td>
<td>100</td>
<td>38.75</td>
<td>39.74</td>
<td>37.80</td>
</tr>
<tr>
<td>K-Fold Cross-Validation</td>
<td>165</td>
<td>20, 5 folds</td>
<td>51.88</td>
<td>64.09</td>
<td>43.58</td>
</tr>
</tbody>
</table>

**Table 5.** Entity evaluation measures for final named entity recognition (NER) model, built using 165 documents, an 80/20 train/test split of the data, and 100 iterations.

<table>
<thead>
<tr>
<th>Entity Class</th>
<th>F1 Score (%)</th>
<th>Precision (%)</th>
<th>Recall (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Animal</td>
<td>79.50</td>
<td>74.42</td>
<td>85.33</td>
</tr>
<tr>
<td>Assay</td>
<td>67.61</td>
<td>72.73</td>
<td>63.16</td>
</tr>
<tr>
<td>Correlates</td>
<td>78.37</td>
<td>79.34</td>
<td>77.42</td>
</tr>
<tr>
<td>Vaccines</td>
<td>55.71</td>
<td>68.42</td>
<td>46.99</td>
</tr>
</tbody>
</table>

**Interactive Visualization**

A total of 98,065 preprints and publications were available on the iSearch COVID-19 Portfolio published between 11/1/2019 and 12/31/2020 at the time of download (Figure 3). The classifier identified 2,409 research articles as relevant to CoPs against SARS-CoV-2. This iSearch dataset was used to populate the dashboard.

**Figure 3.** The final project workflow in which preprint and published documents from the OPA iSearch COVID-19 Portfolio were run through the customized machine learning classifier and named entity recognition models.

Currently, the Tableau dashboard has three visualizations: a summary table, class comparison charts, and a class comparison treemap. The summary table includes the 2,409 research articles from the iSearch dataset identified as relevant by the classifier and related data. In this table, users can filter on publication date, publication type, source of publication, title, first author, last author, author affiliation, entity class, and relevance score. Each paper is marked “Yes” if it contains any of the entity classes. Hyperlinks are also provided to allow users access to the full text of each paper at its original source or in PubMed. Users can also view any clinical trials related to publications via hyperlinks to ClinicalTrials.gov (Figure 4). The class comparison charts provide a visualization of the total instances of each type of entity within each class and allows for filtering on any entity types by clicking on a specific bar, as well as all the filters available to the summary table (Figure 5). The class comparison treemap is a visualization similar to the class comparison charts, but with the treemap format and a corresponding bar chart. (screenshot not shown).
The visualized results are currently available for internal use only and, as such, are housed behind NIAID’s firewall. Public access may be available in the future.

**Discussion**

With the rapid growth of COVID-19 literature, keeping up to date on emerging research for topics such as the CoPs to SARS-CoV-2 has been challenging. In this pilot project, we implemented open-source machine learning methods to build a classifier to identify literature relevant to CoPs against SARS-CoV-2, constructed a custom NER model to extract entities, and visualized the results using a dashboard.

To address the influx of COVID-19 literature, many resources and NLP solutions have been or are being developed to organize this information and address high priority scientific questions within the medical and research...
communities. For example, the Semantic Scholar team at the Allen Institute for AI and partners have provided CORD-19, a free resource of more than 280,000 scholarly articles on SARS-CoV-2. The Semantic Scholar team have also created several NLP tools to explore the large corpus of COVID-19 literature, such as SPIKE-CORD that extracts important information and SciSight which visualizes associations between concepts appearing in the literature. The National Center for Advancing Translational Sciences at the NIH has developed the OpenData Portal, an interactive dashboard that provides SARS-CoV-2 datasets and associated manuscripts in real-time as well as user-friendly heatmap visualizations that permit direct comparison of compounds and repurposed drugs across multiple different assays. Efforts to either cluster or categorize similar articles include the Allen Institute for AI’s SPECTER and LitCovid’s assignment of broad categories (i.e. Treatment and Mechanism) to relevant articles. Additional groups have applied existing NER models trained on biomedical or clinical text while others have created their own NER models. For example, Wang et al. used the CORD-19 corpus to create a model called CORD-NER that covers 75 fine-grained entity types. However, none of these resources and tools are specific to CoPs against SARS-CoV-2, highlighting the need for additional tools and visualizations that are more domain-specific such as our own.

For both the classification and NER tasks, we initially considered only COVID-19 related publications for training a machine learning model. However, we did not accrue a large enough corpus to train and test machine learning models. The lack of relevant research articles was likely related to the status of vaccines at the time of our search in October 2020 but may also have been affected by the initial exclusion of preprints from our dataset. To address our small dataset, we extended our search for papers of CoPs back 20 years to include the first SARS-CoV outbreak in 2002 and the MERS outbreak in 2012. By including research articles of other coronaviruses, we hoped to capture CoPs that may be pertinent to SARS-CoV-2. In addition to these coronaviruses, we identified relevant research articles from different infectious diseases such as HIV, influenza, Ebola, malaria, and measles. The diversity of infectious diseases within our dataset may have provided a wider variety of CoPs compared to a SARS-CoV-2-only dataset.

Even though there are existing tools like LitSuggest to classify and retrieve relevant articles, we proceeded to train our own classifier using scikit-learn. After training several models on our COVID-19 + non-COVID-19 dataset, we tested our selected model, a L2 regularized linear model with stochastic gradient descent learning, on our COVID-19 dataset. Overall, the model had an accuracy of 90.1%. However, it is important to note that part of this dataset was used to train the model. We reviewed the documents of the 42 false positives and found most of these documents initially had differing decisions regarding relevance between the two annotators. Despite the low precision (56.7%), we were more focused on the recall (100%) since our main intent was to ensure we did not exclude any relevant papers.

Our NER model did reasonably well with an overall F1 score of 74.01%. Performance of the NER model, especially of entity classes like vaccines, may be improved by adding more documents that contain these entities. As described above, the number of manually labeled entities was heavily skewed towards CoPs. We did observe that entity classes like vaccines could pose a challenge for any NER model. The format of vaccine names may not allow for vaccine entities to be distinguishable from other entities that use a combination of hyphens and alphanumeric characters, such as “ChAdOx1 nCoV-19”, a vaccine name, compared to “SARS-CoV-2 nucleoprotein ELISpot”, an assay. With the latest version of spaCy v3.0 and the addition of transformer-based features, we may observe additional gains in the performance of the NER model.

Currently, the data feeding the interactive visualization tool is static. Addressing this limitation is important since the Tableau dashboard does not display any new findings regarding CoPs after the publication date 12/31/2020. We are working on a solution with the OCICB at NIAID to make the visualization dynamic. Part of the solution will be to download the latest research on a daily and weekly basis from either the iSearch COVID-19 Portfolio or another source such as CORD-19. Another limitation of our pipeline is that we applied the NER model to only the titles and abstracts of relevant articles. It is very likely that we missed other CoPs or other entities that would have been described further in the full text of the paper. Lastly, our pipeline does not provide context as to whether the identified CoPs were positively or negatively correlated with protection against SARS-CoV-2. This knowledge is important for determining the true protective value of CoPs, such as CD4+ T cells, since it not yet conclusive in the literature which CoPs have an effect.

Despite some of these limitations, our project allows the user to view all research related to the topic of CoPs. We provide several options to not only perform keyword searches but also filter the data to specific articles based on publication date, author name, author affiliation, publication source, publication type, entity class, and entity category.
For example, a user can filter to the companies of three vaccines currently authorized for use in the United States, Pfizer, Moderna, and Janssen, by using author affiliation. With the entity class comparison charts, a user can observe co-occurrence of entities within the literature and explore any similarities and differences among entities (ex. CoPs by vaccine type). We hope that this visualization of data extracted from biomedical literature can provide key insights and trends in CoPs against SARS-CoV-2 until there are consensus definitions for CoPs in the scientific community. The Tableau dashboard may help to fill this information gap while human CoPs data accumulate to benefit vaccine developers, regulatory agencies, and ultimately the public.

**Conclusion**

In this pilot project we developed a user-friendly visualization tool for individuals to stay up to date on the latest literature related to the CoPs against SARS-CoV-2 as well as give individuals key scientific insights from emerging research. To create this dashboard, we employed machine learning and NLP methods with open-source software. We identified research articles relevant to CoPs against SARS-CoV-2 with a machine learning classifier and extracted terms of interest within the titles and abstracts via a customized NER model. Our next steps are to make further improvements to the visualizations and add new features based on user feedback. In addition, we plan to automate the update process of the dashboard by pulling in new publications on a daily or weekly basis.

**Funding**

SLF and SJ were supported by an appointment to the National Institute of Allergy and Infectious Diseases (NIAID) Emerging Leaders in Data Science Research Participation Program. This program is administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy (DOE) and NIAID. ORISE is managed by ORAU under DOE contract number DE-SC0014664.

All opinions expressed in this paper are the authors’ and do not necessarily reflect the policies and views of NIAID, DOE, or ORAU/ORISE.

**Acknowledgements**

We would like to thank Alexis Allot and Qingyu Chen of the U.S. National Library of Medicine, NIH for their expert suggestions in developing the machine learning classifier; Firat Tiryaki of the School of Biomedical Informatics, University of Texas Health Science Center at Houston for his technical assistance with the NER IAA calculations; Jared Kern, Tanvir Ahmed, and Brock Smith of the Analytics, Visualizations, Insights, & Data Science team within the OCICB at NIAID, NIH for their help with the development and launch of the dashboard; Mark Rustad of the Office of Data Science and Emerging Technologies (ODSET) in NIAID, NIH for his improvements to the dashboard; and Steve Tsang of the ODSET in NIAID, NIH for his insights and advice throughout this project.

This work utilized the computational resources of the NIH HPC Biowulf cluster as well as the OCICB HPC cluster at NIAID, Rockville, MD.

**References**

Identifying Opioid Use Disorder from Longitudinal Healthcare Data using a Multi-stream Transformer

Sajjad Fouladvand, MSc1,2, Jeffery Talbert, PhD1,3, Linda P. Dwoskin, PhD4, Heather Bush, PhD5, Amy Lynn Meadows, MD6, Lars E. Peterson, MD, PhD7,8, Steve K. Roggenkamp, MSc1, Ramakanth Kavuluru, PhD1,2,3, Jin Chen, PhD1,2,3
1Institute for Biomedical Informatics; 2Department of Computer Science; 3Department of Internal Medicine; 4Department of Pharmaceutical Sciences; 5Department of Biostatistics; 6Department of Psychiatry; 7Department of Family and Community Medicine, University of Kentucky, Lexington, KY, USA; 8American Board of Family Medicine, Lexington, KY, USA;

Abstract

Opioid Use Disorder (OUD) is a public health crisis costing the US billions of dollars annually in healthcare, lost workplace productivity, and crime. Analyzing longitudinal healthcare data is critical in addressing many real-world problems in healthcare. Leveraging the real-world longitudinal healthcare data, we propose a novel multi-stream transformer model called MUPOD for OUD identification. MUPOD is designed to simultaneously analyze multiple types of healthcare data streams, such as medications and diagnoses, by attending to segments within and across these data streams. Our model tested on the data from 392,492 patients with long-term back pain problems showed significantly better performance than the traditional models and recently developed deep learning models.

1 Introduction

Early identification and engagement of individuals at risk of developing an opioid use disorder (OUD) is a critical unmet need in healthcare1,2. Individuals with OUD often do not seek treatment or have internalized stigma about OUD that limits identification through traditional means, such as screening and clinical interview3. Significant disparities limit access to treatment for OUD resulting in less than 20% of all individuals with OUD receiving any form of treatment in the past year4. While there are currently tools developed to predict aberrant behavior when prescribing opioids5 or to predict OUD from a general primary care population7, there are only a few clinical tools, such as the Opioid Risk Tool5, developed for assessing the risk of OUD. Typical clinician workflow does not allow for comprehensive OUD screening, but available administrative and clinical data have the potential to help clinicians identify and screen higher risk patients providing an opportunity for primary care professionals to play a greater role in increasing OUD detection, treatment, and prevention. Healthcare data are a growing source of information that can be harnessed together with machine learning to advance our understanding of factors that increase the propensity for developing OUDs as well as those that aid in the treatment of the disorders8,9. In healthcare data, patients’ outcomes and treatments are collected at multiple follow-up times. Tools developed to analyze longitudinal healthcare data and to extract meaningful patterns from these ever growing data are critical in addressing real-world public health emergency including but not limited to OUD.

Analyzing real-world data is a complicated task with multiple computational challenges including high dimensionality, heterogeneity, temporal dependency, sparsity, and irregularity10. In particular, healthcare (and claim) data are typically collected from multiple sources, and the subsequent data analysis requires simultaneous analysis of the temporal correlation among multiple streams such as medications, diagnoses, and procedures. Deep learning models have demonstrated great potential in addressing some of these challenges and creating promising longitudinal healthcare data analysis tools. Among them, Doctor AI11, RETAIN12, and DeepCare13 modeled multiple data streams including medications, diagnoses, and procedures using Recurrent Neural Network (RNN) models such as Long-Short Term Memory models (LSTMs)14. Doctor AI concatenated multi-hot input vectors to predict subsequent visit events11. RETAIN used two separated RNNs to generate attentions at the visit level and the variable level as well12. These applications demonstrate that RNNs are promising in longitudinal and sequential healthcare data analysis, since RNNs are capable of extracting contextual information from past time steps and pass this information forward; this helps to efficiently model long-term dependencies in input streams15. Nevertheless, the network architecture and design preclude RNNs from processing long streams in a reasonable amount of time16. Attention mechanism was introduced in RNNs to increase their capacity in capturing long range dependencies more efficiently16-18. Attention-based models
bridge the gap between different states in RNNs using a context vector. Successful applications of multiple attention layers led to the transformer model\textsuperscript{19}, which removed recurrence in RNNs relying entirely on the attention mechanism.

The transformer is a type of attention-based deep learning models originally proposed for natural language processing (NLP) tasks such as machine translation\textsuperscript{19}. Later, transformers have been applied on longitudinal EHR data\textsuperscript{20} to predict patients’ outcomes in the future. There are already several models that have been successfully applied on EHR data without significantly changing the network architecture or loss\textsuperscript{21–23}. Of course, the typical transformer's structure can be altered to better fit the special needs of solving healthcare problems\textsuperscript{20, 24}. Choi et al. proposed a transformer model for healthcare data analysis by utilizing the conditional probabilities calculated from the encounter records to guide the self-attention mechanism in the transformer\textsuperscript{20}. BEHRT\textsuperscript{24} was developed based on BERT\textsuperscript{25}, a popular transformer model for NLP tasks, for analyzing EHR data. BEHRT considers the patients’ existing diagnoses and demographic data to predict their future diagnoses. Similar to RNNs, transformers have been modified to model multiple data streams. Li et al developed a two-stream transformer to analyze both time-over-channel and channel-over-time features in human activity recognition tasks\textsuperscript{26}. Two parallel, yet separate transformers were used to handle two input streams. Another multi-stream transformer has been developed to generate effective self-attentions for speech recognition\textsuperscript{27}. They parallelized multiple self-attention encoders to process different input speech frames. Gomez et al. developed a multi-channel transformer for sign language translation using one self-attention encoder\textsuperscript{28}. Their model finds the attentions across three different channels, i.e. hand shapes, mouthing, and upper body pose. A more recent work\textsuperscript{29} showed that “transformer is all you need” by using multiple transformer encoders. The encoded outputs can be concatenated using a joint decoder that enables simultaneous model training. There are also works that analyze multi-stream data using transformer by simply stacking or parallelizing multiple transformer models\textsuperscript{30, 31}.

Although the recently developed transformer models showed promising performance, especially on handling multiple data streams, the potential of applying transformers on healthcare data analysis has not yet been fully explored. One of the major limitations is the lack of capacity to model multiple data streams within the self-attention layer. The transformer was originally designed to process one data stream, which is mostly an order of words in a NLP task, at a time. The modified transformers either can only handle multiple streams at intra-stream level or they are not suitable to solve OUD identification problem as a real-time task where only previous clinical events can be used to make a decision at a specific time point. Here, OUD identification is a complex data analysis task that includes not only finding long term effects of prescription opioids such as morphine and fentanyl, history of diagnoses such as mood disorders, but also the hidden associations between patient’s prescriptions and diagnoses, since these input streams are highly correlated with each other. Identifying the relationships within and between input streams may reveal hidden patterns leading to an increased classification ability and interpretability for OUDs. Moreover, the medication application patterns and the interactions between medications across different visits as well as the patient’s diagnoses patterns throughout his/her medical history may carry important information that should be extracted in order to develop precise and sensitive OUD identification tools.

This study proposes a novel transformer model called Multi-stream Transformer for Predicting Opioid Use Disorder (MUPOD) to analyze longitudinal healthcare data collected from multiple sources and predict the onset of OUD. First of all, MUPOD is capable of analyzing multiple data streams, such as medication and diagnosis, simultaneously and

\textbf{Figure 1:} Data preprocessing and patient representation. EHR data are first converted to an enrollee-time matrix $X(P, T, F)$. Then, the data are fed to LSTM models to encode the medication and diagnosis streams separately.
extracting associations within and between the streams. Second, MUPOD utilizes attention weights within and across data streams to interpret the classification results. In our experiment, MUPOD successfully captured the complex associations within and between multiple streams including medications, diagnoses, and demographic information, and predicts the onset of OUD precisely.

2 Materials and Methods

Data Set

The large-scale administrative records in the IBM (formerly Truven Health Analytics) MarketScan Commercial Claims\textsuperscript{32} database were used to train and test both baseline models and MUPOD. Data include person-specific clinical utilization, expenditures and enrollment across inpatient, outpatient, prescription drug and carve-out services. The database contains about 30 million enrollees, a nationally representative sample of the US population with respect to sex (50% female), regional distribution, and age.

We extracted medications, diagnoses and demographic information of 682,402 patients who have at least one diagnosis of OUD (ICD-9: 304.0x, 305.5x and ICD-10: F11.xxx; where x can be any code) from 2009 to 2018. The hypergeometric\textsuperscript{33} test was used to identify sub-cohorts of OUD with high statistical significance of whether a population consists the richest information of OUD. We identified an OUD sub-cohort (p-value 0.00) with 229,214 patients who had at least one Clinical Classification Software code (CCS) of 205 (patients with Spondylosis; intervertebral disc disorders; other back problems). This sub-cohort was defined as the case cohort. Note that CCS 205 has already been shown to be a prevalent diagnosis in OUD patients in the literature\textsuperscript{34, 35}.

The case cohort (OUD positive and CCS 205) was matched with a subpopulation of OUD-negative patients called the control cohort. All the individuals in the control cohort have the same back pain diagnosis (CCS 205) but do not develop OUD. We first matched cases and controls based on age and sex. Second, we matched them based on the opioid medication use duration. Specifically, we grouped every opioid medication with a therapeutic class generic product identifier (TCGPI) of 65x as opioid medications. Buprenorphine and Methadone were excluded as they are often used as a treatment for opioid overdose. Next, we randomly sampled OUD-negative patients who have the matched age and gender with the case ensuring that the averaged opioid use ratio between case and control is almost equal.

Table 1 shows the characteristics of cases and controls regarding age, sex, top-10 most frequent medications and top-10 most frequent diagnoses. The diagnoses and the medications were classified using CCS codes and Generic Product Identifier codes (TCGPI) respectively. We grouped opioid analgesics, anticonvulsants (neuromuscular agents), musculoskeletal therapy agents, and antianxiety agents based on the first two digits of their TCGPI codes as 65x, 72x, 75x, and 57x, respectively. The rest of medications were classified using the first 6 digits of their TCGPI codes (from left to right). The variables presented in Table 1 have already been reported as OUD risk factors in the literature\textsuperscript{34, 36}. Especially, diseases including “Other connective tissue disease”, “Other nervous system disorders”, “Essential hypertension”, “Mood disorders”, “Other non-traumatic joint disorders and Anxiety disorders” have been found to be more prevalent diagnoses among OUD patients than normal people\textsuperscript{34}. Note, since we matched the case and control cohorts based on age, sex and analgesics-opioid use, these three variables have similar statistical characteristics across both case and control cohorts. However, the distributions of other variables vary across the case and control cohorts and can be utilized by our deep learning models to discriminate OUD-positive patients from OUD-negative individuals.

Data Pre-processing

For each of the enrollees in the case and control cohort, his/her medications and diagnoses between Jan 2009 and Dec 2018 and demographic records were extracted. In total, we extracted 78,136,935 medication records and 143,275,864 diagnoses records. The original format of the prescription and professional service encounter claims in IBM MarketScan data is a table where each row is a visit and columns are enrollee ID, date of visit, and prescription/diagnoses. If an enrollee has multiple visits, each visit will occupy a row in the table. To facilitate further study of the temporal patterns in the data, we converted the data into an enrollee-time matrix $X(P, T, F)$ where each $x_{i,j} \subseteq F$ is a set of medications or diagnoses (from feature space $F$) associated with enrollee $p_i \in P$ at time slot $t_j \in T$, where $P$ is the
Table 1: Distributions of age, sex, medication, and diagnoses in case and control patients. Top 10 diagnoses and medications are provided. The numbers indicate the number of patients who had at least one such diagnosis or medication.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Case</th>
<th>Control</th>
<th>Variables</th>
<th>Case</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Demographics</strong></td>
<td></td>
<td></td>
<td><strong>Female (percentage)</strong></td>
<td>109,121 (55.60%)</td>
<td>117,699 (59.98%)</td>
</tr>
<tr>
<td>Age (SD)</td>
<td>45.62 (13.81)</td>
<td>52.35 (14.39)</td>
<td>165,112 (84.14%)</td>
<td>180,250 (97.47%)</td>
<td></td>
</tr>
<tr>
<td>Sex</td>
<td>109,121 (55.60%)</td>
<td>117,699 (59.98%)</td>
<td>165,112 (84.14%)</td>
<td>180,250 (97.47%)</td>
<td></td>
</tr>
<tr>
<td><strong>Diagnoses (CCS Code)</strong></td>
<td></td>
<td></td>
<td><strong>Medications (TCGPI Code)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other connective tissue disease (211)</td>
<td>152,703 (77.81%)</td>
<td>165,112 (84.14%)</td>
<td>Analgesics - Opioid (65)</td>
<td>190,141 (96.89%)</td>
<td>196,246 (100%)</td>
</tr>
<tr>
<td>Other nervous system disorders (95)</td>
<td>138,866 (70.76%)</td>
<td>141,350 (72.03%)</td>
<td>Neuromuscular Agents Anticonvulsants (72)</td>
<td>105,508 (53.76%)</td>
<td>97,444 (49.65%)</td>
</tr>
<tr>
<td>Essential hypertension (98)</td>
<td>106,299 (54.17%)</td>
<td>132,049 (67.29%)</td>
<td>Musculoskeletal Therapy Agents (75)</td>
<td>106,186 (54.11%)</td>
<td>102,888 (52.43%)</td>
</tr>
<tr>
<td>Mood disorders (657)</td>
<td>97,035 (49.45%)</td>
<td>81,306 (41.43%)</td>
<td>Antianxiety Agents (57)</td>
<td>76,830 (39.15%)</td>
<td>75,463 (38.45%)</td>
</tr>
<tr>
<td>Other aftercare (257)</td>
<td>127,131 (64.78%)</td>
<td>133,920 (68.24%)</td>
<td>Proton Pump Inhibitors (492700)</td>
<td>71,243 (36.30%)</td>
<td>86,561 (44.11%)</td>
</tr>
<tr>
<td>Residual codes; unclassified (259)</td>
<td>136,177 (69.39%)</td>
<td>152,748 (77.83%)</td>
<td>Serotonin-norepinephrine Reuptake Inhibitors (581800)</td>
<td>58,039 (29.57%)</td>
<td>48,323 (24.62%)</td>
</tr>
<tr>
<td>Other non-traumatic joint disorders (204)</td>
<td>134,042 (68.30%)</td>
<td>150,660 (76.77%)</td>
<td>Selective Serotonin Reuptake Inhibitors (581600)</td>
<td>69,665 (35.50%)</td>
<td>65,005 (33.12%)</td>
</tr>
<tr>
<td>Anxiety disorders (651)</td>
<td>91,736 (46.75%)</td>
<td>78,296 (39.90%)</td>
<td>Hmg Coa Reductase Inhibitors (394000)</td>
<td>53,806 (23.93%)</td>
<td>79,201 (40.36%)</td>
</tr>
<tr>
<td>Disorders of lipid metabolism (53)</td>
<td>94,507 (48.16%)</td>
<td>122,322 (62.33%)</td>
<td>Non-barbiturate Hypnotics (602040)</td>
<td>46,965 (23.93%)</td>
<td>44,404 (22.63%)</td>
</tr>
<tr>
<td>Medical examination/evaluation (256)</td>
<td>129,224 (65.85%)</td>
<td>147,268 (75.04%)</td>
<td>Nonsteroidal inflammatory (661000)</td>
<td>87,301 (44.49%)</td>
<td>98,639 (50.26%)</td>
</tr>
</tbody>
</table>

enrollee set and $T$ is the set of monthly slots between Jan 2009 and Dec 2018. We excluded patients from $X(P, T, F)$ if the number of valid entries is less than 3.

The goal of data representation is to learn a function: $f_R : X \rightarrow \mathbb{R}^d$, where $d$ is 10 in this work and it shows the dimension of the representation to which each input stream is mapped, $X \in \{M, D\}$, and $M$ and $D$ are medication and diagnosis, respectively. To train the function $f_R$, LSTM was adopted. The outputs from all LSTM hidden states were used to represent both the OUD case and control cohorts. The general schema of the data pre-processing and representation is shown in Figure 1.

**MUPOD Architecture**

MUPOD is a transformer-based deep learning model designed to analyze $n$ highly correlated healthcare data streams simultaneously. To minimize ambiguity, the algorithm is described for a single patient and for $n = 3$. Each patient can be represented by $p = (S, y)$ in which $S$ is a set of input streams and $y$ is the target label. Herein, three input streams are considered: 1) medication tuples $(T, M)$ in which $t_i$ is the $i^{th}$ time step and $M$ is a list of medications that the patient is prescribed with at time $t_i$, 2) diagnoses tuples $(T, D)$ where $t_i$ is the $i^{th}$ time step and $D$ is a list of
diagnoses assigned to the patient $P$ at time $t_i$, 3) demographic tuples $(T, G)$ in which $t_i$ is the $i^{th}$ time step and $G$ is the demographic information of patient $P$ at $t_i$.

This study uses the encoder part of transformer to identify the associations between medication and diagnosis across time and detect the onset of OUD. Medications $M$, diagnoses $D$, and demographics $G$ are fed to the model in parallel. The first step is to incorporate the temporal patterns of the data stream into the encoder’s inputs using positional encoding. The embedding layer in the transformer is replaced by the proposed LSTM based representation layer. This change has two computational advantages. Firstly, it deals with challenges in the input data such as variable dimension and data sparseness, which is common in longitudinal healthcare data. Secondly, it extracts hidden parameters and transforms the original input into a new feature space where cases and controls are better separated than in the original feature space.

The encoded input streams are plugged into the attention layer to generate Query, Key, and Value matrices for each input stream. For example, medications $M$ are fed to a set of fully connected layers to generate $M_Q$, $M_K$, and $M_V$, representing the query, key, and value matrices for the encoded medication stream for patient $P$. Let $X, Y \in \{M, D\}$, the Query, Key, and Value matrices are used to find the attentions across these three input streams:

$$Attention(X_Q, Y_K, Y_V) = \text{softmax}(\frac{X_QY_K^TY_V}{\sqrt{d_k}})$$

Note, the $d_k$ is the same as the original transformer. Figure 2a describes how the data flows through the different layers of MUPOD. The raw medication and diagnose streams are first represented in the representation layer (the intermediate outputs of the LSTMS models in Figure 1). The temporal information is then encoded into the represented streams in the temporal encoding layer. The encoded streams are processed in the MUPOD’s multi-stream encoder layer. This novel multi-attention layer is further described in more details in Figure 2b. In the figure, $X_Q$, $X_K$, and $X_V$ represent query, key, and value matrices for stream $X$ ($X \in \{M, D\}$). All possible combinations of the streams are used to determine the attention weights between different visits and across streams. Attentions are then passed through a set of dense layers to generate outputs. For example, given two data streams $M$ and $D$, we can generate three combinations i.e. $MM$, $MD$, and $DD$. 

---

**Figure 2**: MUPOD architecture. $X_Q$, $X_K$, and $X_V$ represent query, key, and value matrices for the input stream $X$, where $X \in \{Medication, Diagnoses\}$. $\text{Att}_{XY}$ represents the attention weights between different records across input streams $X$ and $Y$, where $X, Y \in \{Medications, Diagnoses\}$. $O_{XY}$ represents the outputs, which capture the associations between the input streams $X$ and $Y$. The demographic information is plugged into the system before the last layer and in the classification layer.
The reconstruction layer receives the relevant outputs and maps them to appropriate format for the next layer as described in Equation 2. For example, only the outputs relevant to the medications \( M \) including \( O_{MM} \) and \( O_{MD} \) are used to reconstruct the medication stream appropriate to be fed into the next encoder layer:

\[
f : O_{XX}, O_{XY} \rightarrow \hat{X}
\]

\[
\hat{X} = [\text{Concat}(O_{XX}, O_{XY})]W_x + b_x
\]

where \( O_{XX}, O_{XY} \subset \{O_{MM}, O_{MD}, O_{DD}\} \), \( \hat{X} \in \{\hat{M}, \hat{D}\} \), \( X, Y \in \{M, D\} \), \( W_x \) and \( b_x \) are trainable reconstruction weight and bias matrices. The two reconstructed matrices generated by the last encoder layer are fed to classification layer to make the final decision for the current patient \( p \) as \( \text{Softmax}([\text{Concat}(\hat{M}, \hat{D})]W + b) \).

### Experimental Results

All the deep learning models in this work were deployed on the TensorFlow platform\(^9\) and were trained using eight GeForce GTX 1080 GPUs. The original transformer model, LSTM models, Linear Regression (LR), Random Forest (RF)\(^{10}\) and Support Vector Machine (SVM)\(^{11}\) were compared with MUPOD as baselines. We used 314,504 samples for training, 38,776 samples for validation and 39,212 for testing the models. All results reported in this paper are on the test set. We optimized all models using a random search policy across hyper-parameters of each model. A grid of hyper-parameters values was set up and 10 random combinations of the hyper-parameters were selected to train the models.

The optimized SVM model uses a RBF kernel function and the optimum value for the parameter \( C \) is 0.0039 in this model. The optimized linear regression model uses the L2 norm with \( C = 0.0625 \) in penalization and the sag algorithm as its solver method. The optimum number of trees in the random forest model is 1600 and the optimum value for maximum number of levels in a tree is 40. For the LSTMs, their learning rates were randomly set to \( 10^n \) where \( n \in \{-2, -3, -4\} \). The batch size was randomly selected from \( \{64, 256, 512\} \) and the number of iterations was randomly selected from \( n \times 10^3 \) where \( n \in \{10, 50, 100, 200\} \). The regularization parameter for LSTM models was randomly selected from \( 10^n \) where \( n \in \{-4, -5, -6\} \). The number of hidden neurons for the LSTMs in the representation layer was fixed to 10; because the outputs of these LSTM models were the inputs to MUPOD and the inputs to our model have to be of a fixed dimension (the dimension of our model in this paper is 20: 10 for medications and 10 for diagnoses stream). However, the number of hidden neurons for the other LSTM model used as a baseline (refer to Table 2) was randomly selected from \( 2^n \) where \( n \in \{3, 4, 5, 6, 7, 8\} \).

Table 2 compares the classification performance of MUPOD with LR, RF, SVM, LSTM and the original Transformer model. We used the same train, validation and test data to train, validate and test all models in Table 2 except for the SVM model. Due to the hardware and time limitation we had to train and test this model using 10,000 randomly selected samples. Note, the LSTM model in Table 2 is trained using medication, diagnosis and demographic data. We concatenated the vectors of medication, diagnosis and demographics in each time step and formed a single vector which was fed to this LSTM model. We dynamically unrolled the LSTM model based on the input sequences' lengths and applied a fully connected layer and an argmax function on the last output of the unrolled LSTM model to make the final decisions. The hyper-parameter search space for this LSTM was the same as explained earlier in this section. We used a randomized 5-fold cross validation to tune LR, RF and SVM models. The LR, RF and SVM were trained on the static data and the LSTMs, transformer and MUPOD were trained on the longitudinal data. To create static data for LR, RF and SVM, the longitudinal data was converted to a new format \( Y(P, L) \), where \( P \) is the complete list of patients, and \( L \) is a vector including aggregated values for all medication, diagnosis and demographic features across time steps (from Jan. 2009 to Dec. 2018). In fact, we counted the frequencies for each medications and diagnoses and concatenated these frequencies with demographic information of the patients to create \( L \). Transformer is the original encoder block of the transformer model\(^{19}\). We concatenated the vectors of medication, diagnosis and demographics and fed them to the original encoder block of the transformer model. Then, a fully connected layer and softmax function were used to perform the final classifications. In Table 2, MUPOD has the highest accuracy (0.775), precision (0.741), F1-score (0.790) and AUC (0.871). These results indicate that our proposed model captures important factors in the medication, diagnosis and demographic data and provides an increased power to detect the development of OUD, while LR, RF, SVM, LSTM and original Transformer appear to miss such factors.
Table 2: Performance of OUD classification using MUPOD compared to RF, SVM, LSTM and original transformer.

<table>
<thead>
<tr>
<th>Model</th>
<th>Acc.</th>
<th>Prec.</th>
<th>Rec.</th>
<th>F1-score</th>
<th>AUC</th>
<th>P@R=.8±0.001</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR</td>
<td>0.638</td>
<td>0.641</td>
<td>0.625</td>
<td>0.633</td>
<td>0.689</td>
<td>0.463</td>
</tr>
<tr>
<td>RF</td>
<td>0.698</td>
<td>0.693</td>
<td>0.710</td>
<td>0.702</td>
<td>0.774</td>
<td>0.449</td>
</tr>
<tr>
<td>SVM</td>
<td>0.569</td>
<td>0.539</td>
<td>0.831</td>
<td>0.654</td>
<td>0.677</td>
<td>0.478</td>
</tr>
<tr>
<td>LSTM</td>
<td>0.693</td>
<td>0.784</td>
<td>0.533</td>
<td>0.635</td>
<td>0.790</td>
<td>0.666</td>
</tr>
<tr>
<td>Transformer</td>
<td>0.708</td>
<td>0.654</td>
<td>0.880</td>
<td>0.751</td>
<td>0.801</td>
<td>0.689</td>
</tr>
<tr>
<td>MUPOD</td>
<td>0.775</td>
<td>0.741</td>
<td>0.847</td>
<td>0.790</td>
<td>0.871</td>
<td>0.771</td>
</tr>
</tbody>
</table>

Table 3: OUD classification results for imbalanced test sets. The \( .xN \) means the number of samples in the OUD-positive cohort are \( 0.x \) times smaller than the number of samples in the OUD-negative cohort.

<table>
<thead>
<tr>
<th>Model</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-score</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( .5N )</td>
<td>( .2N )</td>
<td>( .1N )</td>
<td>( .5N )</td>
</tr>
<tr>
<td>RF</td>
<td>.531</td>
<td>.313</td>
<td>.182</td>
<td>.710</td>
</tr>
<tr>
<td>LSTM</td>
<td>.539</td>
<td>.312</td>
<td>.189</td>
<td>.548</td>
</tr>
<tr>
<td>Transformer</td>
<td>.486</td>
<td>.276</td>
<td>.160</td>
<td>.879</td>
</tr>
<tr>
<td>MUPOD</td>
<td>.588</td>
<td>.364</td>
<td>.221</td>
<td>.845</td>
</tr>
</tbody>
</table>

In addition, we tested the models' performances on three imbalanced test data sets with the ratio of OUD-positive samples to OUD-negative samples set to 0.1, 0.2 and 0.5. OUD is an uncommon event and the ratio of OUD-positive to OUD-negative patients in patients who have used Opioid prescriptions at least 3 times is 3.2% in the data set. Therefore, we conducted the experiments in Table 3 to simulate the performance of the models on imbalanced datasets as well. Table 3 shows the model performances on imbalanced test sets. Table 3 shows that MUPOD maintains higher performance on all imbalanced test sets compared to all baselines in terms of precision, F1-score and AUC. Note, we did not show accuracy in Table 3, because this measure is not informative when assessing algorithms on imbalanced data.

We examined the relationships between the medication and diagnosis streams by aggregating the attention weights in the first layer of the model for all the records of each individual and visualized the results. While it is still unclear whether attentions can be used to explain deep learning models\[^{33,44}\] attention weights have been used extensively to assess feature importance\[^{24,42}\]. In particular, the aggregated attentions across all the records of the same patient may be useful to identify important relationships between his/her prescriptions and diagnoses. In the visualization, a rectangular node represents a medication type and an oval node represents a diagnosis code. We divided the accumulated attention weights to “moderate” and “strong” based on pre-defined thresholds (i.e. moderate: \( 0.3 \sim 0.6 \), and strong: \( \geq 0.6 \)) that were selected by visually inspecting the distribution of accumulated attention weights. The moderate and strong connections are represented using dashed and solid lines respectively. The lines of an OUD-negative patient are colored black, while the lines of an OUD-positive patient are colored red.

Figure 3b shows the attention weights computed with MUPOD on one OUD-positive and one OUD-negative patient. The cosine similarities of the medication and diagnosis streams of the two patients are 0.85 and 0.27, respectively, indicating that they have different diagnoses but similar medication records. The connections belonging to the positive and negative patients are well separated. Besides, almost all the strong connections are from the OUD-positive patient, while all the moderate connections are from the OUD-negative patient. Similarly, Figure 3c shows the attention weights on one OUD-positive and one OUD-negative patient. The cosine similarities of the medication and diagnosis streams of the two patients are 0.71 and 0.93, respectively, indicating that they have very similar diagnoses and medication records. Although they have similar records and similar connections between medication and diagnoses nodes, the strengths of attention are different for the OUD-positive patient versus the OUD-negative patient and MUPOD was able to correctly classify these two samples. Note that ONTJD and Opioid are collected with both the OUD-positive
(a) Medication and diagnoses abbreviations and full names.

**Figure 3:** Attention weights. Rectangular nodes represent medications and oval nodes represent diagnoses. Solid, dashed and dotted edges respectively mean strong, moderate and weak connections. We used abbreviations for medications and diagnoses, and provided the full names in (a).

(b) A pair of OUD-positive and OUD-negative samples that have different diagnoses but similar medication records.

(c) A pair of OUD-positive and OUD-negative samples that have very similar diagnoses and medication records.

link (red) and the OUD-negative link (black), indicating the ONTJD-Opioid is often observed on both cases. Figure 3 shows that the attention weights in MUPOD can be used to: 1) discriminate OUD-positive from OUD-negative patients and 2) reveal the relationships between medications that the patient has been prescribed with and the diagnoses he/she has been diagnosed with. These attention weights can further be accumulated across all patients in the cohort to create more generalized conclusions and OUD risk factor identification.

**Conclusion**

OUD is a public health crisis costing the US billions of dollars annually in healthcare, lost workplace productivity, and crime. In this study, we developed a multi-stream transformer model to analyze the long-term impact of medication application pattern, diagnosis history and demographic information, and to explore the associations within and between these streams of patients' data. Our proposed model was able to predict the onset of OUD more effectively compared to baseline models including RF, SVM, LSTM and original transformer model. We discovered that the associations between medication and diagnosis streams are key factors that improve power to predict the development of subsequent OUD.

There are some limitations in our approach. First, the current model relies on patient demographic information and limited subset of medications and diagnoses as features. Incorporating more detailed diagnostic and medication information such as daily dose of opioid could refine the relationship between medications and diagnoses, and create more accurate OUD identification tools. Furthermore, this work only considered a cohort of 196,246 OUD patients who has been diagnosed with the OUD ICD9 or ICD10 codes at least once, ignoring all the undiagnosed OUD patients. For example, more than 224K patients in Truven have been prescribed with Buprenorphine or Methadone but without having any OUD diagnoses. These patients may be undiagnosed OUD patients and could be included in our future work. Second, the current approach cannot predict/estimate risks because the medication application patterns and the diagnosis history of patients that may lead to the increment of OUD risk has not been studied. Third, the explainability of MUPOD was explored using a few representative samples. However, more analysis and correlation analysis using more sophisticated methods such as heatmaps are needed in the future to interpret the model more efficiently. In the future, we will extend our model to address the aforementioned problems such as incorporating more medication and diagnosis features as well as the Morphine Milligram Equivalent (MME) information in MUPOD. The rationale is, given a patient who is constantly on the same type of medication for a while, the variation of the dosage may indicate whether the medication is still effective for the patient.

Despite the limitations of the model, the current approach adds detail to our understanding of the factors that may be important to the development of OUD. Our hope is that a more thorough understanding of the relationships between medications and diagnosis will eventually enable clinicians to identify individuals at risk for OUD at an earlier stage,
and ideally, perhaps even prevent OUD.

Acknowledgments

This research is supported by Kentucky Lung Cancer Research (grant no.KLCR-3048113817).

References


Outcome Prediction from Behaviour Change Intervention Evaluations using a Combination of Node and Word Embedding

Debasis Ganguly¹, Martin Gleize¹, Yufang Hou¹, Charles Jochim¹, Francesca Bonin¹, Alessandra Pascale¹, Pierrho Tommasi¹, Pol Mac Aonghusa¹, Robert West², Marie Johnston¹, Mike Kelly³, Susan Michie²

¹IBM Research Europe, Dublin, Ireland, ²University College London, UK, ³University of Cambridge, UK, ⁴University of Aberdeen, UK

Abstract

Findings from randomized controlled trials (RCTs) of behaviour change interventions encode much of our knowledge on intervention efficacy under defined conditions. Predicting outcomes of novel interventions in novel conditions can be challenging, as can predicting differences in outcomes between different interventions or different conditions. To predict outcomes from RCTs, we propose a generic framework of combining the information from two sources - i) the instances (comprised of surrounding text and their numeric values) of relevant attributes, namely the intervention, setting and population characteristics of a study, and ii) abstract representation of the categories of these attributes themselves. We demonstrate that this way of encoding both the information about an attribute and its value when used as an embedding layer within a standard deep sequence modeling setup improves the outcome prediction effectiveness.

1 Introduction

Randomized controlled trials (RCTs) act as a key source of information about intervention outcomes. An RCT in behavioural science usually captures information on the demographic characteristics of each cohort group in the study, the interventions at a broad level defining a general configuration of each cohort, and a configuration for measuring the outcome values, i.e., some measure of the success for each cohort. A predictive model learned from a set of existing literature could potentially find applications in predicting what is likely to happen for new combinations of cohort groups characteristics, interventions, and outcome measurement settings, which could then provide useful insights to facilitate the process of systematic review. In addition to help compiling systematic reviews, a predictive model may potentially be useful for policy makers to help prescribe a set of behavioural policies that are likely to be helpful to trigger a behaviour change for societal benefits of a target group of people with a given set of characteristics.

Our Contributions. The objective of the paper is to investigate how effectively can the outcome of a behaviour change RCT be modeled in terms of its characteristics comprising mainly the population settings (the whom), interventions (the what) and outcome measurement criteria (the how). We emphasize that the novelty of the paper is not to develop a new neural end-end architecture for the RCT outcome prediction task, for which we employ an end-to-end neural architecture comprising of bidirectional LSTMs, a model that has been met with considerable success in sequence problems such as those of Natural Language Processing.

The novelty of our work rather lies in enriching the embedded input vectors of an end-to-end neural model with additional useful information, the source of which, in our problem, is the ontology of behaviour change attributes, which in addition to the text features proves effective in improving the downstream task of modeling the outcome of an RCT. In particular, our predictive model relies on a novel approach of leveraging information from two sources, namely the annotated text and a document-level co-occurrence relationship between the entities in a behavioural science ontology.

2 Related Work

The work related to text mining for RCTs spans domains from Natural Language Processing to medical informatics. Much of this literature begins with information extraction, which can then be used for summarization, automating (parts of) systematic reviews, or prediction. Early work on extraction from RCTs looked at elements from PICO, and initially classified sentences according to that framework. Instead, a corpus was annotated with PICO entities to

*The study actually uses PIBOSO, which is an extension of PICO.*
The study similarly extract entities but use a much larger inventory of entities taken from the Behaviour Change Intervention Ontology. This and other work has been undertaken to help in automating systematic reviews of RCTs1, 2, 8, 15, which rely on accurate information extraction.

Not much research has yet been carried out on predictive tasks on RCTs. Related to our use case of behaviour change and smoking cessation, the articles16, 17 showed the feasibility of regression approaches to predict the percentage of quitters but this work does not extend to the number of papers and entities that we cover. Probably the closest work to ours is from18. Like our second task of pairwise classification, they look to infer the findings of an RCT based on its intervention, comparator, and outcome entities.

Among existing work that combines text and graph embeddings, joint embeddings of text and relations was employed for link prediction in a knowledge base19. More similar to our embeddings, the authors of20 learn embeddings over a co-occurrence graph of entities and compare them to word embeddings, but they do not explore how those can be combined. A graph-based framework was proposed in21 to incorporate non-local co-occurrences in modeling the semantics between words. While their objective was to improve the effectiveness of word embedding with the help of additional relationships between terms, the objective of our work is to model the relationships between the features in our data with the help of an ontology.

A popular approach towards predictive tasks, such as relation prediction on entities is to use graph convolutional networks22. However, a graph convolution network is suitable for scenarios when an individual instance is modeled as a graph23. In our case, the relations are defined at the level of features and not at the level of each RCT instance. Hence, we leverage the relational information between the features only during the pre-training phase24, so as to generate an enriched set of input vectors to help improve an end-end neural model.

3 Problem Formulation

Ontology Overview. For this study we use the Behaviour Change Intervention Ontology (BCIO) comprised of hundreds of entities at multiple levels of classification25. Lower-level entities, being more granular, define the features used in our study. Different from25, this paper is not concerned with automated extraction of the values of these entities from the papers, but rather assuming that such values have been extracted, to predict the outcome behaviour values and estimated effects given these values.

An RCT as a set of entity-value pairs. A randomized control trial (RCT) study on behaviour science in our dataset usually contains multiple study arms. Each study arm forms an instance for classification and is associated with an outcome value. We represent each document \( d \in \mathcal{D} \) as a set of entity-value pairs. More specifically, an input document \( d \) is a set of 2-tuples of the form \((a, x_a)\), where \( a \) is one of the entities from the BCIO and \( x_a \) is the value associated with the entity, i.e., \( d = \{(a, x_a) : a \in A, x_a \neq \emptyset\} \). The cardinality of the set \( d \) is the number of different entities for which there exists an annotated value. In addition to the semantic type of an entity, each attribute is also associated with a value-type which is also a part of the ontology, and is one of categorical, numerical, or text.

The value of an entity \( a \) (i.e., \( x_a \)), corresponding to an RCT arm, is annotated by a human expert by highlighting the span of text. In our predictive approach, we consider the string (text span) corresponding to each entity value, generally speaking, as a multi-set (bag) of words. The values of each entity are encoded differently depending on the detected type of their annotated span (text, numeric or categorical). For example, each word of a text value is converted to its embedded representation, whereas a numerical token or a categorical value is appended as an additional dimension to a dense vector input (we will revisit this later in the section on outcome value prediction).

Discretizing RCT outcomes. We assume that in each arm, the relationship between the outcome value and the set of features is given by a function of the form \( y(d) = \phi(d) \), where \( y(d) \in [0, 100] \) denotes the percentage outcome value. For the sake of readability, from here we refer to a study arm of an RCT as a document (denoted by \( d \)) in a collection of such arms (denoted as \( \mathcal{D} \)).

https://github.com/HumanBehaviourChangeProject/ontologies
https://www.ebi.ac.uk/ols/ontologies/bcio
The two types of prediction tasks we address correspond to those of a) predicting a discrete interval or range of outcome values, and b) predicting the relative comparison between two studies. Both these tasks require transforming the real valued $y(d)$’s into discrete ones. First, we split the range of $y(d)$, i.e., $[0, 100]$ into a number of intervals. We set this number to 7 to achieve a reasonable degree of discriminability.

We use discrete ranges for the outcome value prediction to simplify the interpretation of the results, e.g., low, moderately low etc. In practice, prediction of a continuous outcome value should be accompanied by a confidence interval, which can be difficult to interpret or even unreliable. Instead of attempting to predict a single exact value, we fix the intervals and try to figure out a relative notion of the likelihood of a low or a high outcome. In our experiments, we also report linear regression results in our experiments, i.e. where the outcome value is directly predicted as a continuous variable.

The start and end-points for each interval constituting a partition of $[0, 100]$, is determined from the distribution of the outcome values in the training set (i.e. the values corresponding to input instances that are known to a model), i.e.,

$$
\mathcal{R}(y(d)) = [0, 100] = \bigcup_{i=1}^{k} [a_i, b_i], \text{ s.t. } a_1 = 0, \quad b_k = 100, \quad a_{i+1} = b_i, \quad Pr[X < a_i] \leq \frac{100i}{k}.
$$

Setting $k=7$ in Equation (1) partitions the range of $y(d)$’s into 7 intervals, where the start of the $i^{th}$ interval is specified by $i^{th}$ 100/7 $\approx$ 14 percentile computed over the distribution of the $y(d)$ values (the percentile points indicating the cut-off points in the cumulative distribution function in Equation (1)). Partitioning the range of the outcome values this way seeks to achieve a uniform binning of the values and mitigate effects of any class priors for the classification task. Each $y(d) \in \mathbb{R}$ is converted to a class label $z(d) \in \mathbb{Z}$ pointing to the index of the interval in which $y(d)$ falls, the intervals being defined as per Equation (1).

**Use-case for Point-wise and Pair-wise Models.** We now describe how the point-wise and the pair-wise models, trained on the input-output associations of existing RCT studies, could potentially be used in practice by an RCT practitioner. In both the point-wise and the pair-wise case, an RCT practitioner would want to know what is likely to happen for a new combination of whom, what, and how features. These features may be entered into a prediction system in the form of attribute value pairs. In the point-wise case, the user would want to obtain a predicted outcome value percentage range on the target population and interventions specified under certain settings independent of a reference point. Instead of intending to obtain a predicted range, for the pair-wise case, an RCT practitioner would want to know if the new combination of input (test) features is likely to increase the success percentage in comparison to a reference study that already exists in the literature.

### 4 Outcome Value Prediction

#### 4.1 Model Overview

Figure [1] shows a standard deep sequence classification model, comprised of stacked layers of LSTMs the hidden states of which lead to a softmax layer of 7 classes into which the outcome value is classified as per Equation (1). This so-called bi-LSTM is a standard neural network architecture that has been met with success in domains that feature an input sequence such as time series or sentences in Natural Language Processing. The novelty lies in defining the scope of the input to this predictive model. More concretely, the input is a set of attribute-value pairs of the form $(a, x_a) \in d$ as annotated in document $d$. The raw input is transformed into a dense vector comprising a pretrained global information about the attribute itself and its text value. Additionally, if the value-type of an attribute is numerical or categorical, its value (a real number or an integer representing the category value) is appended as an additional dimension to the concatenated vector representation comprised of a) the attribute relation and b) the text information. Formally, we denote the transformation from a set of attribute-value pairs to that of dense vectors as

$$
\psi : a, x_a \mapsto \mathbb{R}^{k_f} \times \mathbb{R}^{k_t} \times \mathbb{R},
$$

where each vector corresponds to two distinct subspaces of sizes $k_f$ and $k_t$ (and an additional one for the numerical value). The first subspace (of dimension $k_f$) corresponds to the relations between the entities, whereas the second one (of dimension $k_t$) corresponds to word semantics. The attribute-value set for each arm of a document, $d$, is transformed
Examples of Input Transformation. To illustrate how the attribute-value pairs annotated in a RCT are transformed into inputs to the network of Figure 1, let us look at the following example annotations from a sample paper on smoking cessation studies from our dataset (the example annotations are also shown as the text highlighted in a document at the bottom-left of Figure 1).

Example 1: For an annotated value (text-span) of ‘(% female) 63.3%’ for the ‘gender’ attribute, after tokenizing the string into ‘female’ and ‘63.3’, we obtain the pretrained vector representation of the word ‘female’. We concatenate the word vector ‘female’ with the node vector representation of the attribute ‘gender’, and append the number ‘63.3’ as an additional dimension.

Example 2: For the sample I attribute with value ‘encouraged to set a new quit date’, either the average is computed over vectors (specifically, pre-trained skipgram on PubMed) for the constituent words ‘encouraged’, ‘to’ etc., or the context vector (specifically with Bio-BERT) is obtained for the entire piece of text. This context vector is then used substituted into the $k_t$ dimensional subspace of Equation 2. The other part (the $k_f$ dimensional subspace) is substituted with the node vector for the attribute ‘goal setting’. The additional dimension for the numerical value in this example is 0 (since no number exists in the annotated text instance).

4.2 Textual Context Vector Representations

In this study we investigate two different ways of obtaining the vector representation of the textual context around an instance of an attribute occurrence. These two methods correspond to exploring different granularity for embedding text, one at the level of words and other at the level of sequences of words. Both these approaches are trained on large volumes of unannotated text. While word2vec learns a set of linear transformation parameters for each word to predict its context, BERT captures term semantics with the help of a transformer architecture trained by arbitrarily masking words from text segments.

In our work, we specifically use pretrained word2vec (skipgram) vectors trained on PubMed abstracts. These pre-trained vectors are of 200 dimensions, i.e., $k_t = 200$ in Equation 2. We used zero vectors used for out-of-vocabulary words (8.9% of our dataset). As the context vector, we used the pre-trained Bio-BERT model. The vocabulary of the Bio-BERT is initialized from the larger BERT model of and then fine-tuned on PubMed abstracts. The dimensionality of the feature vectors for the Bio-BERT model is 768, i.e., $k_f = 768$ in Equation 2.

https://bio.nlplab.org/
Embedding of context text is potentially useful to semantically associate/dissociate instances of different/same feature types (e.g., to discover that while two different interventions can be semantically related, the values for the two instances of the same intervention attribute may in fact be semantically different from one another). Next, we describe how we obtain the vector representations of the entities.

4.3 Learning Node Representations

Motivation. One of the limitations of modeling the outcome value as a (predicted) function value of the set of input feature values (comprising numerical, categorical and text features) is that the predictions are likely to be less effective for a sparse feature space. In the context of our problem, sparsity of the feature values is caused due to a wide range of different population characteristics, or interventions used in the studies, e.g., some studies report the average age of a cohort, while others use median age. Moreover, a predictive model assumes that the features are independent. However, in the context of our domain of behavioural science reports, correlations do exist between the entities. For example, if some interventions are likely to work well on a cohort of young people (with lower values for minimum age), they are also likely to work well on cohorts with lower mean age. As another example, sets of interventions are also correlated with each other, e.g., intuitively speaking, ‘psychological counseling’ often works well with ‘continuous monitoring’.

Embedding nodes as vectors has been reported to improve downstream prediction tasks for the biomedical domain, such as modeling interactions between genes, diseases and drugs. In our case, via embedding nodes as vectors we intend to model the correlations between features of different types. For the purpose of graph construction we group the BCIO entities into broader types for who (P) what (I) and how (Q).

Graph Construction. The first step in our proposed approach is to construct an undirected graph \( G = (V, E) \) intending to capture the co-occurrences between different feature instances. Each node in this graph is represented by a tuple \( v(t, a, x_a) \), \( a \) being an attribute of type \( t \in \{P, I, Q\} \). Formally,

\[
V = \{(t, a, x_a) : \exists (a, x_a) \in \mathbf{x}_d, \ d \in \mathcal{D} \},
\]

where \( t \in \{P, I, Q\} \) and the node set, \( V \), is thus comprised of nodes of unique types with unique values. While constructing a node corresponding to an attribute, only its categorical or numerical value is included as a part of the node. We exclude the string (text span) of the annotation for an attribute because including it would make the graph too fine-grained (e.g., one node for each possible value of an intervention ‘goal setting’). This would lead to sparse edge relations between its nodes, which in turn would not be conducive for modeling the inter-attribute relations. Note that the text information is eventually used in the downstream prediction task because it constitutes a separate subspace of the input vectors (Equation 2). Next, we define the edges in \( G \) as follows. Formally, an edge exists between a pair of nodes corresponding to the values of entities of type \( t \) and of type \( t' \) (\( t, t' \in \{P, I, Q\} \) \( t \neq t' \)), if these values are observed in the same RCT arm (document). To model the likelihood of a correlation between pairs of attribute-values, we set the weight of an edge \( e \), \( w(e) \), to reflect the relative number of times such associations between the feature values are observed across a number of different documents in the collection.

Node Embedding. After constructing the graph \( G = (V, E) \) from a given collection of documents \( \mathcal{D} \), the next step in our proposed method is to obtain a dense vector representation for each node of \( G \). Specifically, we applied the random walk based node2vec algorithm to learn the vector representation of each node. The choice of visiting a next node in node2vec is controlled by two parameters, namely a) the (inverse) return parameter, \( p \), which if set to a low value makes it more likely for the walk to return to \( t \), and b) the in-out parameter, \( q \), which if set to a high value makes the walk unlikely to visit nodes that are not adjacent to \( t \). A low value of \( p \) and a high value of \( q \) is thus likely to make the walk more compact. Specifically, for our experiments, we tie the two parameters by setting \( q = 1 - p \).

In our case, after applying node2vec on the weighted graph of Equation 3, attribute-value combinations that are likely to be correlated to each other will be embedded close to each other, because these nodes are likely to be more reachable from each other with a random walk using the edge weights as probabilities.
5 Modeling RCT Comparisons

We now extend the point-wise prediction framework to learn a comparison function between a pair of RCTs. A practical use-case for this pairwise situation arises when an RCT practitioner wants to compare a new combination of population, intervention and outcome settings with reference to an existing study, which we call the reference study involving a target population. The intention of the predictive model in this case is predict if a new combination of interventions is likely to increase the success ratio in comparison to another study, which is different from predicting if the relative comparisons between two arms of the same study yield significant differences.

For pairwise modeling of RCTs, the input is a pair of RCTs. The attribute-value pairs of each RCT is transformed to a variable length sequence of embedded representations of concatenated node and word vectors identical to the input transformation of Equation 2 (Figure 1). The pairwise prediction model employs a Siamese type architecture, where we feed in a pair of RCTs as input. The training phase makes use of the annotated attribute-value pairs of existing RCTs reported in the literature. The encoded representation of the LSTM layer for both studies is then concatenated before applying a sigmoid layer. During training, the ground-truth label between a pair of RCTs is 1 if the outcome value of the first element of the pair is less than that of the second, or 0 otherwise. The network is trained with all combinations of pairs of the form $(d_1, d_2)$ from the training set. A pair $(d_1, d_2)$ is used in the training set in a unique ordering, i.e., inclusion of $(d_1, d_2)$ excludes $(d_2, d_1)$, which means that the total number of pairs used for training is $|D|(|D| - 1)/2$. In the testing set, one of the elements of each pair is a new combination of PIQ features, unseen in the training set, whereas the other is from the training set (i.e., a previously seen reference study).

6 Evaluation

Dataset. For our experiments, we focused on the domain of the smoking cessation behaviour change RCTs reports the compilation of such a dataset (called HBCP) of behaviour change RCTs focused on smoking cessation; however, their dataset is mainly targeted towards addressing information extraction from RCTs. Since we focus on a different task, that of predicting outcomes of RCTs, we construct an extended version of the HBCP dataset for our experiments. Different for our classification task the RCT instances are constructed as described in Equation 2. Each study includes a number of study arms corresponding to a fundamental unit of a study, i.e., a particular population group with certain characteristics and a set of interventions applied on the group. Outcome values are reported separately for each arm and a single RCT can have multiple arms.

Our extended dataset comprises a set of 513 RCTs (PDF documents) on behaviour change for smoking cessation. The annotation schema of our dataset follows the ontology and the guidelines defined in . A team of in-house domain experts annotated a total of 7451 attributes of different types from the set of 513 PDF documents. Table 1 presents an overview of our dataset.

### Table 1: Dataset Characteristics

<table>
<thead>
<tr>
<th>#Docs</th>
<th>513</th>
<th>#Arms</th>
<th>1064</th>
<th>#P (whom) attributes</th>
<th>4808</th>
<th>#I (what) attributes</th>
<th>5129</th>
<th>#Q (how) attributes</th>
<th>2554</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean y(d) (Outcome %)</td>
<td>16.8</td>
<td>Median y(d) (Outcome %)</td>
<td>13.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table 2: Summary of the best results with 5-fold cross validation for the point-wise (7-class outcome classification and regression) and the pairwise tasks. The first five are ablation baselines.

<table>
<thead>
<tr>
<th>Embedding Method</th>
<th>Point-wise</th>
<th></th>
<th>Pairwise</th>
</tr>
</thead>
<tbody>
<tr>
<td>None Values-Only</td>
<td>0.5532</td>
<td>15.05</td>
<td>0.6237</td>
</tr>
<tr>
<td>Skipgram Text-Only</td>
<td>0.6344</td>
<td>10.11</td>
<td>0.7350</td>
</tr>
<tr>
<td>Skipgram Text+N2V-1Hot</td>
<td>0.6456</td>
<td>7.47</td>
<td>0.7282</td>
</tr>
<tr>
<td>Bio-BERT Text-Only</td>
<td>0.6745</td>
<td>7.33</td>
<td>0.7479</td>
</tr>
<tr>
<td>Bio-BERT Text+N2V-1Hot</td>
<td>0.6946</td>
<td>7.52</td>
<td>0.7429</td>
</tr>
<tr>
<td>Skipgram Text+N2V</td>
<td>0.6658</td>
<td>8.04</td>
<td>0.7585</td>
</tr>
</tbody>
</table>
| Bio-BERT Text+N2V | **0.7072** | **7.06** | 0.7553 **|**
**Figure 2:** Parameter sensitivity effects of Text+N2V (Bio-BERT) for point-wise outcome value classification for different context sizes. It can be observed augmenting pre-trained feature relationship information as a part of the input produces substantially better results in comparison to the Text-Only and the Text+N2V-1Hot approaches (shown as the two constant lines).

**Figure 3:** Parameter sensitivity measured in terms of RMSE (lower the better) of Text+N2V (with Bio-BERT) for point-wise outcome value regression for different context sizes. A comparison with Figure 2 shows that regression results are more sensitive to parameter variation effects.

**Setup.** To assess the effectiveness of the graph-based approach for the point-wise and the pairwise prediction tasks, we compare our proposed approach of joint embedded input representation (text and attribute-value nodes) with two ablation baselines\(^*\). As the first ablation baseline, we employ a standard one-hot encoding of each attribute node coupled only with its numeric value (i.e., \(k_f\) being number of unique attributes, \(k_t = 0\) in Equation 2), which is equivalent to standard linear regression and multi-class classification (for the continuous or the interval prediction). Note that this baseline neither uses information from the text around the context of the attribute instances, nor does it use an embedded representation of the attributes themselves. We name this baseline ‘Values-only’.

The second ablation baseline, Text-Only, does not use any information from the attribute-value pair co-occurrences, i.e., we feed in as input vectors to the network of Figure 1 (and its pairwise equivalent) an aggregation (average) of word vectors, each of dimension \(k_t\) (see Equation 2), from the annotated text spans. For a numeric value, e.g., mean age of a population, we feed in its value as an additional dimension in the input vector along with the word vector representation of its context. We used two different ways of obtaining the feature vectors for the text, namely a) Skipgram, where we used pre-trained skipgram vectors of dimension 200 (i.e. for this baseline \(k_t = 200\)), and b) Bio-BERT, where we used the pre-trained Bio-BERT model\(^*\) to obtain \(k_t = 768\) dimensional representation of the context text.

The third ablation baseline, Text+N2V-1Hot, employs a one-hot encoding of the attribute-values nodes (Equation 3). This baseline treats each graph node as independent ignoring the co-occurrence relations between the edges. The two different text embedding approaches lead to two different settings for the one-hot experiments with different feature dimensions (for the text part).

To test the approaches, we employ 5-fold cross-validation. The intervals to induce the class labels are computed on each training fold instance (Equation 1). For pairwise classification, training proceeds with pairs from the training fold. The test instances are constructed by pairing up each RCT from the test fold with each from the train fold, the objective being to predict if a new study, for which the outcome is not known, is likely to yield a higher or a lower outcome compared with an existing one.

---

\(^*\) Implementation of the point-wise and pairwise models, along with the dataset would be made publicly available. [https://huggingface.co/emilyalsentzer/Bio_ClinicalBERT](https://huggingface.co/emilyalsentzer/Bio_ClinicalBERT)
Table 2 summarizes the best results obtained with each method for the point-wise and the pairwise tasks. We observe that a value-only based approach (similar to a simple linear regression or a multi-class classification) produces not too effective results. We observe from the bottom part of Table 2 (‘Text+N2V’) that leveraging information from the co-occurrence likelihoods between the behaviour science attributes in the form of embedded node representations improves significantly (t-test with 95% confidence) the effectiveness of both the point-wise and the pairwise tasks in comparison to the corresponding text-only approaches (e.g. compare the ‘Skipgram Text-Only’ results with ‘Skipgram Text+N2V’ ones). Moreover, the results also improve in comparison to the approach when the node attribute features are treated as independent one-hot vectors (e.g. compare ‘Bio-BERT Text+N2V’ results with ‘Bio-BERT Text+N2V-1Hot’ ones). The pairwise case yields slightly better results when skipgram vectors are combined with the node embeddings.

Parameter Sensitivity. In addition to presenting the best results for each method in Table 2, we now investigate the effects of varying the parameters of node2vec for obtaining the embedded vectors that are concatenated as inputs to the architecture of Figure 1, i.e., parameters - the context size (ws), dimension of embedding (kf), and the return/in-out node2vec parameters (p, q). We explore the parameter space only for the most effective combination method of Table 2 i.e. the ‘Text+N2V’ with the Bio-BERT embedding.

Figures 4 and 5 report parameter sensitivity for the multi-class classification and regression tasks, respectively. From Figure 4, we observe that smaller values of p (and thereby larger values of q = 1 − p) usually result in better outcome value prediction. As per Equation 4, small values of p (and large values of q) are likely to yield locally compact walks. In the context of our problem, this means too much exploration on the co-occurrence graph may introduce noise in the form of false long-chain dependencies across entity values of different types.

For the pairwise case, we explore the parameter space for the combination of node vectors with skipgram vectors (since this configuration produced better results than the Bio-BERT ones). Figure 4 shows trends that are similar in nature to that of Figure 2 i.e., the optimal results are obtained for smaller values of p.

### 6.1 Prediction with Uncertainties

In this section, we investigate the feasibility of a more pragmatic approach where only a small subset of the documents in a collection is annotated with the attribute-value information. This scenario also tests how effectively can a prediction system, trained on a subset of the collection (called the seed set), may subsequently be used to make predictions for newly created research articles on behaviour change (i.e., those for which no manual annotations are available).

For each unannotated documents, we employed the unsupervised information extraction method to automatically extract a set of attribute-value pairs, given its text as an input to the extractor. The prediction system is then trained on a mixture of both manually annotated (hence, clean) and automatically extracted (hence, uncertain) data.

To conduct experiments for predictions with uncertainties, from our static collection of annotated documents, we first use only a fraction of the data as the seed set (signal), and then employ the extractor to automatically infer the attribute values from the remaining set (noise). Figure 5 shows how does the effectiveness of our prediction model (Text + N2V) is affected by the use of automatically extracted values (the RMSE values are averaged over 5-fold CV test splits). The red line plots the RMSE values obtained only with the seed data, whereas the blue line, for each fraction of the seed
data, displays the results obtained by augmenting the seed data with extracted information from the remaining fraction of the data.

It is seen that too small or too large a seed set (i.e. \( \leq 10\% \) or \( \geq 60\% \)), the use of additional uncertain data, in the form of automatically extracted attribute values, is not able to outperform the results obtained with clean data only. However, it is seen that using about 20\% of clean data, coupled with 80\% additional (potentially noisy) data improves the overall outcome value prediction effectiveness. This implies that knowledge gained from new RCTs in the form of extracted attribute-value pairs can potentially be injected into our prediction system for improving its effectiveness.

7 Conclusions

We investigated how effectively can we automatically predict outcomes from RCTs on behaviour change studies. The novelty lies in encoding an RCT instance as a combined representation of the embedded textual context of annotated values coupled with the embedded representation of the relations between attribute-value instances. Our experiments demonstrate that this way of modeling the inputs outperforms the cases which make an oversimplifying assumption that such attribute-value instances are independent. A broader impact of our work is that it shows that the outcome value of a behaviour change study can be predicted within satisfactory levels of accuracy, which implies that AI systems can potentially be used by policy-makers in implementing a set of behaviour change policies (interventions) on a target population.

In future, we would like to investigate outcome prediction for RCTs with automatically extracted attribute values from documents.

References


Impact of COVID-19 Pandemic on Emergency Department Visits: A Regional Case Study of Informatics Challenges and Opportunities

Hamid Ghaderi1*, Jeffrey R. Stowell1,2*, Murtaza Akhter1,2, Craig Norquist3, Paul Pugsley1,2, Vignesh Subbian1
1The University of Arizona, AZ, United States; 2Valleywise Health, AZ, United States; 3Honor Health, AZ, United States

Abstract

In this paper, we examined informatics challenges and opportunities related to emergency department visit data during public health emergencies. We investigated the impact of COVID-19 pandemic on the volume and acuity of adult patients visiting the emergency department (ED) of a medical center in Arizona during the pandemic compared to the pre-pandemic period. We performed a negative binomial regression analysis to understand how different public health-related mandates and statewide business opening/closing orders in Arizona affected the daily emergency department visits. The results of this study show that the average daily ED visits decreased by 20% during the COVID-19 pandemic in comparison with the same period in 2019. In addition, the business closure order had the most impact on emergency department visits in comparison to other public health mandates.

Introduction

As the COVID-19 pandemic developed and intensified throughout the United States in the early part of 2020, state governments and healthcare systems enacted a range of mitigation strategies and operational changes to address the increasing number of infections in the community. Initial public health strategies recommended avoiding unnecessary healthcare utilization to decrease virus spread and to ensure that there is enough capacity to handle spikes in COVID-19 cases [1]. In addition, different public health-related mandates, including the stay-at-home order, business closures, and mandatory wearing masks were implemented to prevent and mitigate the spread of the virus [2]. While the number of patients who visited Emergency Departments (EDs) due to COVID-19 significantly increased, health systems noticed a change in ED visits for acute care unrelated to COVID-19 [3]. Accordingly, the pattern of ED visits changed with the spread of COVID-19 in 2020. In this regard, collecting, sharing, and analyzing ED volume data are necessary for improved coordination of care and distribution of resources during high utilization times. An example of using available ED resources equitably was the Boston Marathon Bombing, where trauma patients were appropriately distributed such that trauma centers in the Boston metropolitan area were not overloaded. By recognizing spikes in patient arrivals during either mass casualty events or public health outbreaks, each hospital could call in backup providers, offload the least sick of the inpatients, and be better prepared for a potential surge of sick patients in a more timely and appropriate manner. Conversely, recognizing drastic decreases in volume ahead of time in a metropolitan area could serve as a warning or alert that something large is happening such as a new wave of pandemic or fear of becoming sick from the healthcare system. However, there are several issues relating to the ED volume data elements that have not been examined in the literature. In addition, to the best of our knowledge, the change in the pattern of ED volume and acuity, as well as the impact of public health-related mandates and state-wide business opening/closing on ED visits due to the COVID-19 pandemic, are yet to be studied in detail while considering regional factors. Therefore, in this study, we seek to answer the following questions.

1) What are the challenges of collecting, coordinating, and sharing ED volume data elements and recommendations to address those challenges?

2) How did the COVID-19 pandemic affect ED visits and acuity (in terms of Emergency Severity Index (ESI) level at triage, admissions, and specific ED activations) in a single regional hospital compared to the pre-pandemic period?

3) How did different periods of the COVID-19 pandemic in Arizona, as defined by different public health-related mandates and statewide business openings/closings, affect daily ED visits in a single regional hospital?

* Equal contribution
Background

The United States declared a national emergency on March 13, 2020, in response to the COVID-19 pandemic. Afterwards, states imposed stay-at-home orders to prevent the spread of COVID-19 and relieve the burden on the healthcare system, which influenced the number and type of ED visits. Several prior studies have examined the impact of the pandemic on ED visits. We provide a review of results from these studies, along with data limitations, and other informatics challenges.

Hartnett et al. [4] investigated the effect of the COVID-19 pandemic on ED visits in the United States from January 1, 2019 to May 30, 2020. They found that ED visits in the United States declined 42% during the pandemic. They also demonstrated that ED visits associated with the patients aged less than 14 years, females, and Northeast region had the sharpest decline after the pandemic. Additionally, they showed that the proportions of infectious disease-related ED visits during the early pandemic period in 2020 were four times higher than those in the same period before the pandemic in 2019. Adjemian et al. [5] presented an update on Hartnett et al.’s [4] report; they compared the effect of the pandemic on ED visits in the United States during December 20, 2020–January 16, 2021 with those during a pre-pandemic period (December 15, 2019–January 11, 2020). According to Adjemian et al.’s [5] report, ED visits were 25% lower in December 2020-January 2021 than in the same months the previous year. In addition, Adjemian et al. [5] showed that, during the pandemic period, higher proportions of ED patients, particularly pediatric patients, are seeking treatment for mental and behavioral health issues.

Although these previous studies on the effect of COVID-19 on ED volume in the United States (i.e., [4] and [5]) show a decline in ED volume during the pandemic, there are life-threatening conditions that always require immediate emergency care, even during the pandemic. In this regard, Lange et al. [6] conducted a study of ED visits for three life-threatening acute health conditions, including myocardial infarction, stroke, hyperglycemic crisis, in the United States immediately before and after the declaration of the COVID-19 pandemic as a national emergency. They reported that ED visits for each of these health conditions declined after the declaration of the COVID-19 pandemic as a national emergency. The US ED volume data included in [4-6] were from National Syndromic Surveillance Program (NSSP) that collaborates with programs across CDC, state, and local health departments to collect electronic health data in real-time. So the findings in these studies are subject to several data-related limitations that are as follows: (1) diagnostic categories rely on the usage of certain codes which may be missing or utilized inconsistently across different institutions; (2) hospitals reporting to NSSP change over time when new facilities are established, and less frequently, when they close; and (3) NSSP data collection coverage is not uniform across all states. That is, in some states nearly all hospitals report their data, while in others, only the hospitals in certain counties report their data.

The decrease in ED volume in the United States during the pandemic could have been caused by the public fear of seeking medical care, causing the patients with time-sensitive critical conditions to delay their treatment. Kim et al. [7] compared the number of patients with serious diagnoses visiting seven EDs in greater Chicago, Illinois, during the pandemic period to that of the pre-pandemic period. This study demonstrated sharp reductions in ED visits for all major diagnosis classes (e.g., cardiac, surgical, orthopedic, neurologic, gastrointestinal, chronic respiratory) during the COVID-19 pandemic. In this study, the data are from a particular healthcare system Chicago, so the results may not be generalizable to other settings with different patient population characteristics and/or local government responses to the COVID-19 pandemic. Also, this study relies on ICD-10 coding data, which are based on the treating physician’s impression. To address this limitation, Kim et al. [7] excluded symptom-based diagnoses and included the diagnoses that often need objective findings obtained during ED evaluation. Although Kim et al. [7] used a valid and publicly available system to link ICD-10 codes to clinical diagnoses, it is possible that not all relevant ICD-10 codes for serious diseases are captured.

Jeffery et al. [1] examined ED visits and visits that led to hospital admissions in healthcare systems in five states, including Colorado, Connecticut, Massachusetts, New York, and North Carolina covering a four-month period before and during the COVID-19 pandemic. This cross-sectional study demonstrated that the ED visits in these five states significantly decreased from 41.5% to 63.5% during the pandemic. They also showed that the rates of the ED visits that led to hospital admissions were stable before the pandemic but increased from 22% to 149% during the pandemic. Since Jeffery et al.’s [1] study data are limited to ED visits to the healthcare systems of only five states, the results may not be generalizable to the other states in the United States. In addition, because their study data did not include diagnoses, they were unable to examine how ED patient case-mix have changed over time.

Westgard et al. [3] conducted a study of ED visits, demographics, characteristics, and diagnoses of the patients visiting an urban Level-1 trauma center in the United States 28 days before and 28 days after declaring a state emergency on March 13th, 2020. The results of their study show that ED visits significantly decreased after the state of emergency declaration. They also demonstrated that the decline in ED visits significantly changes by patient demographics (i.e., age, gender, and race) and visit characteristics (i.e., insurance status, arrival mode, and disposition). The data in this study are limited to only one center and a short study period, limiting the generalizability of the findings to other medical centers and the rest of the pandemic period.
Butt et al. [8] studied the volume and acuity of the patients visiting four EDs affiliated with general, cardiac, cancer, and obstetrics hospitals in Qatar during the pre-pandemic and pandemic periods. They demonstrated that the ED patient visits in both general and specialty hospitals substantially reduced during the pandemic. They also showed that the cardiac hospital and cancer hospital experienced the sharpest decline and the least decline in ED visits, respectively. Furthermore, the proportions of severely ill patients visiting EDs sharply dropped during the pandemic. The included EDs account for more than 80% of all ED visits in Qatar, making the results generalizable to all EDs in Qatar. The data in this study are limited to the first two months after the diagnosis of the first COVID-19 patient in Qatar, so the findings may change by increasing the length of the study time frame. In addition, the study data are limited to only the volume and acuity of the ED patients before and during the pandemic. Although the results show a significant decrease in ED volume due to the pandemic, the travel restrictions imposed by the state of Qatar due to the COVID-19 pandemic may have affected the volume and acuity of ED patients. However, the data studied by Butt et al. [8] did not include travel restrictions, so they could not examine the impact of travel restrictions on ED visits.

Compared to adults, children had a lower risk of COVID-19 infection, and therefore needed less emergency medical care [9]. Consequently, although many studies focused on the volume, epidemiological, and clinical characteristics of the ED visits related to the adult patients with COVID-19, data and studies on pediatric patients are limited. With this in mind, Lozzi et al. [10] studied the pediatric ED visits of five hospitals across Italy from March 1 to March 27, 2020 and showed that pediatric ED visits in March 2020 decreased by 73% in comparison to the same period in 2019. The data in this study are from only one pediatric ED in Italy, reducing the generalizability of the findings. Similar to Italy, pediatric ED visits in other countries changed during the pandemic. According to CDC, routine pediatric care in the United States decreased during the pandemic [11]. Additionally, an examination of Chinese outpatient pediatric care showed a drop in overall outpatient healthcare as well as a significant decrease in infectious complaints [12].

Social distancing, on the other hand, posed different risks to children. For example, data from a large pediatric hospital showed a threefold increase in dog bites, as well as a significant increase in national calls to poison control centers, including a high percentage of child exposures. Chaiyachati et al. [13] studied the effect of social distancing and a state-wide stay-at-home order on pediatric ED visits to a tertiary urban children’s hospital in the United States during the COVID-19 pandemic. This study compared the pediatric ED visits in 2020 after an official stay-at-home order with the ED visits in the last three years from 2017 through 2019 and showed reduced visits, increased acuity, and consistent chief complaints in 2020 in comparison with the last three years. Given that chief complaint scripts are not standardized across hospitals, the generalizability of this study is limited.

The review of literature shows that quite a few studies have been done to investigate the impact of the COVID-19 pandemic on the volume and acuity of ED patients. However, the impact of different public health mandates on ED visits; and the challenges and opportunities associated with ED visit data during the pandemic have not been extensively studied in the literature, which we seek to address in this study.

Methods

Data source

The dataset used in this study comes from a single medical center in Arizona and was collected from January 1, 2019 to August 31, 2020. The data elements were sourced from electronic medical records, quality dashboards that report the daily and monthly census, and additional specialty reports to track ESI and admission locations. Quality metrics were obtained from the hospital quality committees that track such data for state and national reporting. Diagnoses categories that are “suspected” in the ED and get validated later in the course of the admission are collected and collated by existing committees as mentioned above. We used quality reporting such as MIDAS, or other validated data sets to extract information regarding ST-Elevation Myocardial Infarction (STEMI). STEMs were obtained from the cardiovascular service line which reviews all arrivals and validates if the case was a STEMI or not for very specific reporting requirements. An overview of the informatics challenges and opportunities for improvement were obtained through direct in-person interviews with the clinical research team.

Common data elements

The initial study metrics abstracted from every ED arrival for 2019 influenza season (pre-pandemic) and 2020 (pandemic) included arrival time, arrival method, disposition type, disposition time, gender, age, ESI, chief complaint, lab results (e.g., flu A, flu B, respiratory syncytial virus, strep, respiratory pathogen polymerase chain reaction, and COVID-19) imaging results (e.g., chest X-ray (anteroposterior/lateral vs portable)). Also, the hospital metrics include daily hospital census, including total daily ED physician staffing hours, total daily trauma activations, total daily codes, transfers from an outside hospital, transfers out of the system, total daily admissions to each location (i.e., Intensive Care Unit (ICU)), acute, and Acute Palliative Care Unit), and status (i.e., total daily observations, and total daily admissions). We revised the dataset based on availability and the likelihood of obtaining data from various EDs in Arizona. The following are the current set of common data elements that we believe are needed to perform ED volume studies:
(1) Daily ED visit metrics: total adult ED visits and total pediatric ED visits;
(2) Monthly ED visits for each of the ESI categories (ESI is a measure of acuity, and each patient should be assigned a number from 1 (highest acuity) to 5 (lowest acuity) i.e., ESI 1, ESI 2, ESI 3, ESI 4, and ESI 5);
(3) Monthly ED admissions which include admissions for medical/surgical floor, step down/intermediate, and ICU;
(4) Monthly quality metrics that include ED stroke activations, ED ST-Elevation Myocardial Infarction (STEMI) activations, ED trauma activations, and ED sepsis activations. It should be noted that our study includes only adult patients, and pediatric patients were not considered in our study.

Data collection and coordination process

Coordination with other sites for obtaining data included reaching out to individuals who we thought would be interested and capable of obtaining data. This largely included previous or ongoing ED-based colleagues. We found that rapidly establishing was challenging and previous agreements did not include such uses of data for publication and would have required soliciting individual agreements with each institution to use the data. Consequently, the dataset used in this study is limited to only one medical center in Arizona, which reduced the generalizability of results. For example, we were unable to provide information regarding alternative institutions, such as urgent care, critical access, or suburban hospitals. Furthermore, given the issues and challenges that we had in obtaining data, we simplified the dataset to improve the likelihood of obtaining data. However, it is obvious that resolving data collection issues and obtaining more detailed information will result in more educated or specific findings and recommendations. For instance, simply knowing the raw volume of patient arrivals is a good starting point, but it is insufficient to know if either an increase or decrease in volume also has an increase or decrease in acuity. Are sicker patients staying home or unable to get care elsewhere? What was the impact of primary care offices closing? Did those patients requiring medication refills for diabetes or heart medications go to urgent care or EDs for refills or did they simply go without and subsequently suffer worse outcomes due to the forced non-compliance with their medications? Other challenges we faced include separating the pediatric population from the adult population since many EDs do not have a separate pediatric department that is open 24 hours per day, and they combine and merge the census. Consequently, the data on ED adult patients had to be manually extracted from the electronic health records.

Challenges and opportunities related to data sharing

Data protection regulations

The majority of data sharing challenges were largely related to privacy and security concerns of Protected Health Information. Local institutions must approve research projects involving human participants, and this process follows internal protocols. Even within the same county, institutional review board (IRB) protocol preparation and approval processes are generally different. The main variations are based on the structure and content of the documentation, the bureaucratic processes necessary for the documentation, the time it will take for the institution to process and approve it, and the number and kind of revisions requested by the IRB. As a result, such differences served as a barrier to multisite research and led to potential delays in starting the studies. These challenges were more pronounced for sites that do not have sufficient research administration capacity.

The use of de-identified data sets or population-level data removes some if not all the concerns regarding possible re-identification of patients, but it also removes some level of necessary detail. For instance, if we wanted to see which zip codes, ethnicities, ages, or genders were most impacted, it would have been more complicated to coordinate how best to protect patients via either a safe harbor or other methods of de-identification that would also allow aggregation amongst others.

Most institutional IRBs are established to review traditional research on drugs or devices and are less familiar with more of a population-based approach nor are they comfortable with sharing data across disparate institutions. Early approaches to obtaining IRB approval as well as data collection were delayed in part due to this misunderstanding. Using more relevant language regarding de-identified data, limited data sets, and quality improvement data may have served us better.

Data formats and report structures

Some healthcare organizations tend to have established departments for the submission of data to government and regulatory agencies in many different formats to comply with the specific needs for each request or requirement. Unfortunately, many of these established departments have become less flexible to modifications of or even sharing of their data collections for fear of manipulation or alteration of it. A report created for reporting to the Centers for Medicare & Medicaid Services is either overly cleaned or processed and often misses the details that physicians expect and need to ‘tune’ their internal diagnostic algorithms. Taken out of context, meaningful cases reported for disease entities such as heart attacks or strokes may seem either exaggerated or minimized, depending on the strict requirements for inclusion or exclusion. Instead of having a readily available source of heart attack, stroke, sepsis, or even ED visit data in an understandable format, we were often met with unclear or even confusing datasets created not to improve patient care but to comply with various uncoordinated agencies. Some data were necessarily delayed
until a diagnosis was assured, either from pathology, autopsy, or discharge diagnosis reports, making real-time reporting more challenging.

Instead of replacing existing departments and reporting structures, we recommend concurrent reporting of the same data in more clinically relevant formats that can better be used for monitoring population movements and clinical outcomes as well as improvement of processes. One example could be reporting of suspected severe sepsis or septic shock versus those diagnosed. By the time the definitive diagnosis is made, it could be hours after the patient has left the ED. It is important to provide feedback that a patient either was or was not septic to better recognize and diagnose the next patient, but often physicians are less concerned about the specific reporting requirements of sepsis versus severe sepsis in individual patients.

**Study design**

We conducted a descriptive, cross-sectional study of volume and acuity of adult patients visiting the ED of a medical center in Arizona in 2020 (the pandemic period) compared to those in 2019 (the pre-pandemic period). We stopped data collection in August 2020, which may limit generalizability through the remainder of the pandemic. We also examined how different public health-related mandates and statewide business openings/closings in Arizona, affected daily ED visits to the medical center. We created scatter plots for comparing the total daily ED visits, monthly ED visits for each ESI (i.e., ESI 1, ESI 2, ESI 3, ESI 4, and ESI 5), monthly admissions to each hospital unit (i.e., medical/surgical unit, step down/intermediate unit, and ICU), and monthly activations of each specialty team (i.e., trauma, STEMI, stroke, and sepsis) in the medical center before the COVID-19 pandemic with those during the pandemic. To compare mean differences from the pre-pandemic period with those of pandemic period, we performed paired Wilcoxon test (also known as Wilcoxon signed-rank test). We created a boxplot and reported summary statistics for four time periods affected by public health mandates and statewide business openings/closings during the pandemic. To compare mean differences from the pre-pandemic period with those of pandemic period, we performed a negative binomial regression to predict the daily ED visits in the following months based on different variables, including being in non-infectious or early infectious periods, potential public health mandates such as business/service closures, stay-at-home order, business/service reopening, and mask mandates. All analyses were performed using R (version 4.0.5) statistical computing software.

**Results**

**Trends in the daily ED visits before and during the COVID-19 pandemic**

Our analyses show that, except for January, daily ED visits during the pandemic period were substantially lower than those during the pre-pandemic period (see Figure 1). The mean (±SD) of daily ED visits during the pandemic was lower than that during the pre-pandemic period (106 ± 25 vs. 132 ± 15, p < 0.05). Indeed, it decreased by 20% during the pandemic. The lowest number of ED visits occurred during April (55 visits). Although the number of daily ED visits during the pandemic has increased since the noticeable decrease in April, it remained lower than the corresponding number of daily ED visits in the previous year.

**Figure 1.** Trends in the medical center’s daily ED visits before and during the COVID-19 pandemic

**Figure 2.** Daily ED visits in four specific periods during the pandemic, including period 1 (early infectious period, Jan 1 - March 29), period 2 (Stay-at-home order, March 30-May 15), period 3 (re-opening to secondary closure, May 16 - Jun 29), period 4 (second closure to current, June 30 – current)
Daily ED visits in four specific periods

As shown in Figure 2, the medical center had its highest daily ED visits (173 visits) during the early infectious period and experienced its lowest daily ED visits (55 visits) following the stay-at-home order. After reopening businesses in Arizona, the daily ED visits increased; then, it went into a small decline following the second closure. However, there is not a significant difference between the mean (±SD) of daily ED visits in the reopening period and that in the second closure period (101 ± 17 vs. 97 ± 12, p = 0.43).

Monthly ED visits for each ESI before and during the COVID-19 pandemic

Every included patient should fit in one of the ESI categories (i.e., ESI 1, ESI 2, ESI 3, ESI 4, and ESI 5). In this measure of acuity, ESI 1 is the highest acuity and ESI 5 is the lowest acuity. ED visits with ESI 1 and ESI 2 are considered high acuity ED visits. The results show that the mean (±SD) of monthly high acuity ED visits during the pandemic was higher than that in the pre-pandemic period (421 ± 46 vs. 363 ± 22, p < 0.05) (see Table 1). Likewise, the proportion of high acuity ED visits during the pandemic was higher than that of the pre-pandemic period (13% vs. 9%, p < 0.05). As evident in Figure 3, the mean (±SD) of monthly ED visits with ESI 3 in 2020 was lower than that in 2019 (2079 ± 394 vs. 2878 ± 246, p < 0.05). However, the mean (±SD) of monthly ED visits with ESI 4 during the pandemic was not significantly different from that during the pre-pandemic period (662 ± 207 vs. 682 ± 51, p = 0.74). The results also show that the mean (±SD) of monthly ED visits with ESI 5 during the pandemic period was lower than that in the pre-pandemic period (26 ± 5 vs. 39 ± 10, p < 0.05). Accordingly, ED visits with ESI 5 had the sharpest decline (33%) during the pandemic in comparison with the other ESI categories, showing the patients with the lowest acuity were most likely to avoid visiting EDs to get medical care during the pandemic.
Figure 3. Trends in the monthly ED visits for each ESI before and during the COVID-19 pandemic. Figure 3 shows that, during the pandemic, the medical center had the lowest number of ED visits for each ESI in April, except for the ED visits with ESI 5.

Table 1. High acuity ED visits

<table>
<thead>
<tr>
<th>Month</th>
<th>Number of ED visits</th>
<th>Number of high acuity ED visits</th>
<th>Proportion of high acuity ED visits</th>
</tr>
</thead>
<tbody>
<tr>
<td>January</td>
<td>4190</td>
<td>4252</td>
<td>393</td>
</tr>
<tr>
<td>February</td>
<td>3681</td>
<td>3875</td>
<td>331</td>
</tr>
<tr>
<td>March</td>
<td>4259</td>
<td>3573</td>
<td>369</td>
</tr>
<tr>
<td>April</td>
<td>4143</td>
<td>2266</td>
<td>373</td>
</tr>
<tr>
<td>May</td>
<td>3937</td>
<td>2655</td>
<td>334</td>
</tr>
<tr>
<td>June</td>
<td>3770</td>
<td>3263</td>
<td>357</td>
</tr>
<tr>
<td>July</td>
<td>4105</td>
<td>3075</td>
<td>384</td>
</tr>
<tr>
<td>August</td>
<td>4024</td>
<td>2905</td>
<td>363</td>
</tr>
</tbody>
</table>

Monthly admissions to hospital care units before and during the COVID-19 pandemic

Patients presenting to emergency departments may get admitted to one of the hospital units (medical/surgical care unit, step down/intermediate care unit, intensive care unit), depending on the level of care needed. The analysis of monthly admissions to each unit shows that the center experienced a sharp decline in the number of admissions to all of its units in April after implementing a business closure order on March 19th and stay at home order on March 30th in the state of Arizona (see Figure 4). In addition, the admissions to each unit of the medical center increased in May following the reopening of businesses on May 16th. Since the declaration of public health emergency on March 11th, the medical center experienced the most admissions to each of its hospital units in June when the businesses were open and mandating wearing masks initiated on June 19th.

According to the results, there is not a significant difference between the mean (±SD) of monthly admissions to the medical center during the pandemic period and that during the pre-pandemic period (690 ± 87 vs. 725 ± 53, p = 0.383). Also, there is not a significant difference between the mean (±SD) of monthly admissions to the step down/intermediate unit during the pandemic period and that during the pre-pandemic period (191 ± 29 vs. 180 ± 21, p = 0.461). The mean (±SD) of monthly admissions to the medical/surgical unit did not significantly change during the pandemic compared with the pre-pandemic period (447 ± 52 vs. 445 ± 39, p = 1). On the other hand, the results show that the mean (±SD) of monthly admission to the ICU during the pandemic is not the same as that during the pre-pandemic period (52 ± 38 vs. 100 ± 10, p < 0.05). Indeed, there is a sharp decline (48%) in the mean of monthly admission to the ICU during the pandemic in comparison with the pre-pandemic period. There is not a significant difference between the proportion of ED visits that get admitted during the pandemic and that during the pre-pandemic period.
period (20.5% vs. 18.2%, p = 0.104). The proportion of the ED visits that become ICU admits during the pandemic is, however, lower than that during the pre-pandemic period (1.5% vs. 2.5%, p < 0.05) (see Table 2).

Monthly activations of each specialty team before and during the COVID-19 pandemic

The analysis of data on the number of activations of each specialty team in the medical center before and during the pandemic showed quite a few different trends. The mean of monthly trauma activations decreased by 15% during the pandemic (105 ± 18 vs. 135 ± 16, p < 0.05). Conversely, the mean of monthly sepsis activations increased by 15% during the pandemic compared to the pre-pandemic period (109 ± 14 vs. 95 ± 9, p < 0.05). There is not a significant difference between the mean (±SD) of monthly STEMI activations during and before the pandemic (2 ± 2 vs. 3 ± 1, p = 0.105). Likewise, there is not a significant difference between the mean (±SD) of monthly stroke activations during the pandemic period and that during the pre-pandemic period (4 ± 1 vs. 4 ± 3, p = 0.573).

Figure 4. Trends in monthly admissions to each unit of the medical center before and during the COVID-19 pandemic. As evident in Figure 4, during the pandemic, the ICU of the medical center had the sharpest decline in the number of admissions compared to the other hospital units.
in the community increased in May voluntary self observed an initial decrease in ED visits started just before these public health mandates, which may be because of of ED visits visiting EDs due to the fear of seeking care in hospitals the number of COVID pandemic period We observed that daily ED visits in comparison with the other public health mandates. On the other hand, accessibility to care is implemented. As public health mandates is considered as a binary variable. A binary variable associated with a specific period is equal to 1 during that period. Also, a binary variable associated with a public health mandate is equal to 1 when that mandate is implemented. As shown in Table 3, business closure, stay-at-home order, and business reopening have a reverse relationship with the daily ED visits and make a decline in it. Also, business closure makes the most decrease in the daily ED visits in comparison with the other public health mandates. On the other hand, according to the results, the mandatory wearing of masks increases the daily ED visits. In addition, the medical center has more ED visits during non-infectious or early infectious periods in comparison with the pandemic period. Furthermore, daily ED visits are more during the early infectious period in comparison with the non-infectious period.

**Table 2. Number and proportion of ED visits that get admitted or become ICU admits**

<table>
<thead>
<tr>
<th>Month</th>
<th>Number of ED visits</th>
<th>Number of ED visits that get admitted</th>
<th>Number of ED visits that become ICU admits</th>
<th>Proportion of ED visits that get admitted</th>
<th>Proportion of ED visits that become ICU admits</th>
</tr>
</thead>
<tbody>
<tr>
<td>January</td>
<td>4190</td>
<td>4252</td>
<td>802</td>
<td>810</td>
<td>118</td>
</tr>
<tr>
<td>February</td>
<td>3681</td>
<td>3875</td>
<td>638</td>
<td>766</td>
<td>84</td>
</tr>
<tr>
<td>March</td>
<td>4259</td>
<td>3573</td>
<td>779</td>
<td>767</td>
<td>97</td>
</tr>
<tr>
<td>April</td>
<td>4143</td>
<td>2266</td>
<td>723</td>
<td>554</td>
<td>105</td>
</tr>
<tr>
<td>May</td>
<td>3937</td>
<td>2655</td>
<td>706</td>
<td>643</td>
<td>92</td>
</tr>
<tr>
<td>June</td>
<td>3770</td>
<td>3263</td>
<td>688</td>
<td>708</td>
<td>102</td>
</tr>
<tr>
<td>July</td>
<td>4105</td>
<td>3075</td>
<td>760</td>
<td>659</td>
<td>96</td>
</tr>
<tr>
<td>August</td>
<td>4024</td>
<td>2905</td>
<td>702</td>
<td>616</td>
<td>103</td>
</tr>
</tbody>
</table>

**Negative binomial regression**

We performed a negative binomial regression to predict the daily ED visits in the next months based on different variables, including being in non-infectious or early infectious periods, potential public health mandates, including business/service closures, stay-at-home order, business/service reopening, and mandatory wearing masks. To do so, the number of ED visits is considered as the dependent variable. Also, each of the specific periods and each of the public health mandates is considered as a binary variable. A binary variable associated with a specific period is equal to 1 during that period. Also, a binary variable associated with a public health mandate is equal to 1 when that mandate is implemented. As shown in Table 3, business closure, stay-at-home order, and business reopening have a reverse relationship with the daily ED visits and make a decline in it. Also, business closure makes the most decrease in the daily ED visits in comparison with the other public health mandates. On the other hand, according to the results, the mandatory wearing of masks increases the daily ED visits. In addition, the medical center has more ED visits during non-infectious or early infectious periods in comparison with the pandemic period. Furthermore, daily ED visits are more during the early infectious period in comparison with the non-infectious period.

**Table 3. Negative binomial regression model results**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Rate Ratio</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>4.721</td>
<td>112.226</td>
<td>99.309 126.852</td>
</tr>
<tr>
<td>Non-infectious period</td>
<td>0.146</td>
<td>1.157</td>
<td>1.023 1.308</td>
</tr>
<tr>
<td>Early infectious period</td>
<td>0.182</td>
<td>1.199</td>
<td>1.064 1.351</td>
</tr>
<tr>
<td>Business closure</td>
<td>-0.349</td>
<td>0.705</td>
<td>0.650 0.765</td>
</tr>
<tr>
<td>Stay-at-home order</td>
<td>-0.011</td>
<td>0.989</td>
<td>0.896 1.093</td>
</tr>
<tr>
<td>Business reopening</td>
<td>-0.157</td>
<td>0.855</td>
<td>0.762 0.958</td>
</tr>
<tr>
<td>Mandatory wearing masks</td>
<td>0.2</td>
<td>1.222</td>
<td>1.122 1.330</td>
</tr>
</tbody>
</table>

**Conclusion**

We observed that daily ED visits during the COVID-19 pandemic substantially decreased compared to the pre-pandemic period. However, it is not clear whether this reduction in the number of ED visits is because of a decline in the number of COVID-19 patients or because of a decline in the number of patients unrelated to COVID-19 who avoid visiting EDs due to the fear of seeking care in hospitals due to the pandemic. Also, the sharpest decline in the number of ED visits was in April 2020 after the implementation of state-wide business closure and stay-at-home orders. We observed an initial decrease in ED visits started just before these public health mandates, which may be because of voluntary self-isolation before the mandatory public health orders. The results show that the number of ED visits increased in May 2020. This could be because of reopening businesses and services, increasing the spread of the virus in the community, and increases in the number of patients who need to visit EDs.
ED visits with ESI 5 had the sharpest decline in comparison with the other ESI categories, suggesting that patients with the lowest acuity were most likely to avoid visiting EDs to get medical care during the pandemic. Although the proportion of ED visits resulting in ICU admission has not substantially changed in comparison with the prior year, the results show a significant decrease in the number of ICU admissions through the ED. This decline in ICU admissions raises concern that critically ill patients may not be getting timely care during the pandemic due to the fear of seeking care or travel restrictions. This ultimately delays interventions for serious health conditions and may also cause a surge in the number of ICU admissions after the pandemic subsides. It is yet to be determined whether there was an actual decrease in the number of patients who need intensive treatment during the pandemic and whether the decline in ICU admissions due to the pandemic increased the mortality rate. If confirmed, suitable strategies should be implemented to provide appropriate and prompt care to critically ill individuals. With this in mind, a more comprehensive dataset from multiple sites and for a longer time period is needed to be able to investigate the effectiveness of the public health mandates. Future extensions of this work include investigating the impact of public health mandates on the daily ED visits while adjusting for seasonality of daily ED visits.

Acknowledgment

This work was supported in part by the State of Arizona Technology and Research Initiative Fund (TRIF).

References

Multi-task deep learning-based survival analysis on the prognosis of late AMD using the longitudinal data in AREDS

Gregory Ghahramani\textsuperscript{1,}\textsuperscript{a}, Matthew Brendel\textsuperscript{1,}\textsuperscript{a}, Mingquan Lin, Ph.D.\textsuperscript{2}, Qingyu Chen, Ph.D.\textsuperscript{3}, Tiarnan Keenan, BM BCH, Ph.D.\textsuperscript{4}, Kun Chen, Ph.D.\textsuperscript{5}, Emily Chew, MD\textsuperscript{4}, Zhiyong Lu, Ph.D.\textsuperscript{3}, Yifan Peng, Ph.D.\textsuperscript{2,}\textsuperscript{b}, Fei Wang, Ph.D.\textsuperscript{2,}\textsuperscript{b}

\textsuperscript{1}Department of Physiology, Biophysics, and Systems Biology, Weill Cornell Medicine, New York, NY USA; \textsuperscript{2}Department of Population Health Sciences, Weill Cornell Medicine, New York, NY USA; \textsuperscript{3}National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM), National Institutes of Health (NIH), Bethesda, MD USA; \textsuperscript{4}National Eye Institute (NEI), National Institutes of Health (NIH), Bethesda, MD USA; \textsuperscript{5}Department of Statistics, University of Connecticut, Storrs, CT USA

Abstract

Age-related macular degeneration (AMD) is the leading cause of vision loss. Some patients experience vision loss over a delayed timeframe, others at a rapid pace. Physicians analyze time-of-visit fundus photographs to predict patient risk of developing late-AMD, the most severe form of AMD. Our study hypothesizes that 1) incorporating historical data improves predictive strength of developing late-AMD and 2) state-of-the-art deep-learning techniques extract more predictive image features than clinicians do. We incorporate longitudinal data from the Age-Related Eye Disease Studies and deep-learning extracted image features in survival settings to predict development of late-AMD. To extract image features, we used multi-task learning frameworks to train convolutional neural networks. Our findings show 1) incorporating longitudinal data improves prediction of late-AMD for clinical standard features, but only the current visit is informative when using complex features and 2) “deep-features” are more informative than clinician derived features. We make codes publicly available at https://github.com/bionlplab/AMD_prognosis_amia2021.

1 Introduction

Age-related macular degeneration (AMD) is the leading cause of vision loss, and is projected to affect approximately 288 million people by 2040\textsuperscript{1,2,3}. In the United States alone, the annual healthcare cost of treating this disease is $4.6 billion, creating an extreme burden on patients and the healthcare system\textsuperscript{4}. AMD is characterized by the destruction of a retinal pigment epithelial (RPE) cells, which directly interact with photoreceptors to allow for proper function of the eye\textsuperscript{5}. In AMD, drusens, or lipid deposits, form near the RPE cells, which can eventually lead to tissue atrophy in the eye. In addition, RPE cells normally contain melanosomes, which create a certain pigmentation in the eye.

The onset of the disease can be heterogeneous between individuals. The majority of patients have a form of the disease known as dry AMD, which has fairly slow progression, whereas some patients (10-15% of early-stage AMD) will develop choroidal neovascularization (CNV) which leads to the rapid loss of vision and faster onset of late-stage AMD\textsuperscript{5}. Therefore, to improve treatment plans for patients, it can be useful to understand the risk of developing CNV and in particular, the risk of developing late-stage AMD.

The current method to assess AMD severity requires the use of color fundus photographs (CFP), which are generated by a low-power microscope that captures general eye health and examines structures within the eye\textsuperscript{5}. These photographs are then sent to grading centers, where experts analyze specific characteristics, including presence, type, and extent of drusens, presence/extent of retinal depigmentation, serous sensory retinal detachments, subretinal hemorrhages, subretinal fibrosis, and geographic atrophy, which are used for characterizing AMD severity. A simplified AMD severity score and risk classification has been developed by the Age-Related Eye Disease Study (AREDS) Research Group\textsuperscript{2}. Based on characteristics from CFPs at the current time of visit, patients are binned into 5 categories (0-4), which estimate the likelihood the patient will progress to late-stage AMD. This five-step simplified severity

\textsuperscript{a}Equal contributions.

\textsuperscript{b}Equal contributions.
scale is the current clinical standard in assessing a patient’s risk of developing late-AMD. This risk is calculated using the size of drusens, presence of pigmentation abnormalities, age, and smoking status at the current time of visit.

Over the past decade, the use of deep learning has grown exponentially. Convolutional neural networks (CNNs) have been used to identify patterns within images to classify medical imaging data. Various models have been developed to characterize CFPs, based on several characteristics. Single task models have been used to classify characteristics such as geographic atrophy and drusen presence. In addition, multi-task models have been developed to characterize these eye characteristics simultaneously. Subsequently, researchers have used these image features derived from a CNN model in a survival setting to predict patients who are at risk of developing late-stage AMD. However, it is well known that the rate of progression for patients within the early-stage AMD category is heterogeneous. In this work we hope to utilize the time-varying information for these patients to improve upon risk-prediction for AMD patients.

We combine elements of multiple past works to improve upon AMD patient stratification, while introducing a novel time-varying component to improve model performance. A multi-task learning model was used to predict both drusen size and presence of pigmentation abnormalities in the right and left eyes of patients. Drusen size and presence of pigmentation abnormalities are the criteria used for the simplified AREDS severity scale. Image features are extracted from the multi-task learning model. Multi-task learning was incorporated to extract more generalizable image features than the clinician derived features, hoping to improve our ability to predict risk of developing late-AMD. Either the image features or clinical features are passed through either a multilayer-perceptron (MLP) or long short-term memory (LSTM) network, to predict patient risk of developing late-stage AMD. A survival loss function is utilized to train the risk prediction model to account for patients that end the study without developing late-stage AMD. We compare model performances of these image derived features with those of expert-derived features to compare our results to a baseline model. The end goal of this work is to (a) reduce the burden on grading centers by reducing the time needed to analyze simple cases and assisting in edge-case classification and (b) use the features derived from the images to improve upon the stratification of patients with early-stage AMD based on the risk of progressing to late-AMD.

2 Materials and Methods

2.1 Dataset

In this study, we use the AREDS cohort sponsored by the National Eye Institute (National Institutes of Health). It was a 12-year multi-center prospective cohort study of the clinical course, prognosis, and risk factors of AMD, as well as a phase III randomized clinical trial to assess the effects of nutritional supplements on AMD progression. In short, 4,757 participants aged 55 to 80 years were recruited between 1992 and 1998 at 11 retinal specialty clinics in the United States. The inclusion criteria were wide, from no AMD in either eye to late AMD in one eye. The AREDS dataset is publicly accessible to researchers by request at dbGAP. In the AREDS cohort, at baseline and at annual visits, comprehensive eye examinations were performed by certified study personnel using a standardized protocol, and CFP (field 2, i.e., 30° imaging field centered at the fovea) were captured by certified technicians using a standardized imaging protocol.

The longitudinal analysis of the AREDS cohort led to the development of the patient-based AREDS Simplified Severity Scale for AMD, based on the grading of CFP. This simplified scale provides convenient risk factors for assessing the risk of progression to late AMD that can be determined by clinical examination or by less demanding photographic procedures than used in the AREDS. The scale combines risk factors from both eyes to generate an overall score for the individual, based on the presence of one or more large drusen (diameter > 125 mm) and/or AMD pigmentary abnormalities at the macula of each eye. The Simplified Severity Scale is clinically useful in that it allows ophthalmologists to predict an individual’s 5-year risk of developing late AMD. This 5-step scale (from score 0 to 4) estimates the 5-year risk of the development of late AMD in at least one eye as 0.4%, 3.1%, 11.8%, 25.9%, and 47.3%, respectively.

In our study, the event of interest was the development of late AMD. The ground truth labels (AREDS Simplified Severity Scale and phenotype features such as drusen status and macular pigmentary abnormalities) used for both training and testing were the grades previously assigned to each CFP by expert human graders at the University of Wisconsin Fundus Photograph Reading Center. The reading center workflow has been described previously.


507
To train and test our models, we created a data subset that consisted of 3,747 patients from AREDS who had not reached late AMD through year 3. Of these patients, 2.7% reached late-AMD by year 5, 6.1% reached late-AMD by year 8, and 9.3% reached late AMD by the end of the study. This dataset consisted of the gradings from the certified study personnel, as well as inferred grades, such as presence of one or more large drusen and AMD pigmentary abnormalities at the macula of each eye.

2.2 Development of the algorithm

Figure 1 shows the overarching architecture used to A) develop the fine-tuned CNN drusen size and pigmentation abnormality classifiers trained on images from all patients and B) extract the "fine-tuned" and "pretrained" features on the images from years 0, 2, and 3 for the patients who had not reached late-AMD by year 3. Figure 1C demonstrates example model of how image features are used to predict risk of developing late-AMD at years 5, 8, and overall risk. We will describe each module in the subsequent sections.

2.2.1 Multi-task Learning

Multi-task learning is a field of machine learning where multiple tasks are learned in parallel while using a shared representation. It exploits the similarities (shared image features) and differences (task-specific image features) between the features present on different tasks, thus reduces the losses of various tasks simultaneously.

In this study, we created a multi-task deep learning model that trains the classification of drusen size and presence of pigmentation abnormalities simultaneously. Drusen size and pigmentation abnormalities are the features used to calculate the 5-step simplified severity scale. Of the 60,929 CFPs used in this study, 41,147 images from visits after year 3 were randomly sampled in a 90-10 split to train and test. Data augmentation was performed to improve model generalizability. Training images were randomly horizontally flipped, cropped, blurred, rotated, sheared, morphed, and the contrast was randomly strengthened or weakened using the ImgAug module in Python. These augmentations were set to plausible realistic ranges of CFPs. All images were resized to 256 × 256, then center cropped to 224×224. This aided in removing the unwanted areas as there is some blank space surrounding the eye image. Images were then normalized to a mean of 0 and standard deviation of 1 based on the mean and standard deviation of 10,000 randomly sampled images from the entire dataset.

Two different models were trained for this study to compare performance. ResNet152 and EfficientNetB3 were used to see how deep learning architecture affects model performance. Weights on both the ResNet152 and EfficientNetB3 models were both initialized to the pretrained ImageNet weights. The last-fully-connected layers of both models were replaced with two separate linear layers, one for predicting size of drusen (three-class) and one for predicting pigmentation abnormality (two-class) (Figure 1A).

---

https://github.com/rwightman/gen-efficientnet-pytorch
To account for the class imbalance, weighted cross entropy loss was used on each classifier. Weightings were set to $\frac{1-p_i}{n-1}$, where $p_i$ is the proportion of class $i$ in the training set, and $n$ is the number of classes. The model loss was set to the mean of the drusen classifier loss and the pigmentation classifier loss. The Adam algorithm was used as an optimizer with the learning rate set to 0.005. Up to 25 epochs, with a batch size of 16 images were run to train each model. Training ceased when the training loss dropped below 85% of the test loss. The weights from the epoch with the lowest test loss were saved and used to extract image features for deep survival analysis.

2.3 Image Feature Extraction

To extract image features, the fully-connected layers from the multi-task learning model were removed, and an n-dimensional vector was extracted from the last hidden layer for each image (2,048 for ResNet152 and 1,536 for EfficientNetB3). These features will be referred to as “fine-tuned” features. A comparison is also done with the ImageNet pretrained model to demonstrate improved feature extraction using the multi-task learning model. Those features will be referred to as “pretrained” features. Collectively, they are referred to as “deep” features.

2.4 Survival Analysis

Deep image features from CFPs at years 0, 2, and 3 and clinical data of 3,297 patients who had not reached late-AMD at year 3 were used to predict risk of developing late-AMD at years 5 and 8 and by the end of the study. Five-fold cross validation was used to evaluate the performance of all models. Of the 80% of the patients not in the test set, 80% were used to train the model and 20% were used as a validation set. One single batch was used during training to account for low numbers of uncensored data.

CoXPH Model Image features from each eye and each visit were concatenated to an $n \times 2 \times m$ vector, where $n$ is the size of the CNN output (2048 for ResNet152 and 1536 for EfficientNetB3) and $m$ is the number of years included. Using principal component analysis (PCA) from the scikit-learn library, a linear transformation to 10 dimensions was made on the training dataset. On average, these PCA decompositions explained 22.4%, 85.4%, 81.6%, and 73.6% of the variance for the ResNet152 pretrained, ResNet152 fine-tuned, EfficientNetB3 pretrained, and EfficientNetB3 fine-tuned feature vectors across the cross-validation training sets, respectively. Fewer than 10 principal components further reduced the explained variance, while more than 10 led to linearity convergence issues with the CoXPH model. The CoXPH Fitter model from the lifelines module was then fit with a step size of 0.1 and no penalizer, using the Breslow method for handling ties.

MLP Model For each patient, deep image features were extracted from images of the patient’s left eye and right eye at years 0, 2, and 3. Features were z-score normalized, fit to the training set, prior to training and evaluating the survival models. The models incorporated time-dependent information in two different ways, both utilizing the pytorch deep learning library. For the MLP model, the left and right eye image features were concatenated for the three different time points, generating a $(n \times 2 \times m)$ dimensional vector. The MLP model consisted of one hidden layer (32-dimension for 1 year and 96-dimension for 3 years) that then fed into a final linear layer with one node as the output for the survival loss calculation.

LSTM For the LSTM model (Figure 1C), the left and right eye data for each time point were concatenated separately for each visit, then passed through a single linear layer to decrease the dimensionality of the concatenated feature vector by a factor of 8, creating a tensor of size $(b \times 3 \times \frac{2 \times n}{8})$, where $b$ is the batch size and $n$ is the size of the feature vector from the CNN. Each of these time points were then used as a separate input into the LSTM to model the time-varying changes in eye features. The hidden state from the final time point was fed into a survival loss similar to what was done for the MLP model. The hidden state size for the model was 128. In addition, dropout ($p = 0.6$) was used prior to the linear layer to reduce overfitting. All models used the pycox loss, which is an approximation of the negative partial log likelihood. We adopted survival loss using Efron’s method to handle ties in the survival time using code from the Pysurvival package.

2.5 Evaluation Metrics

Survival Analysis To evaluate model performance, several different metrics were used. Concordance index was calculated using the pycox package. In addition, patients were categorized into late-stage or not late-stage AMD at

---

1https://github.com/CamDavidsonPilon/lifelines
2 years (year 5) and 5 years (year 8) after the year 3 time point. We then generated area under the receiver operating characteristic curve (ROC AUC) for predicting late-stage AMD at years 5 and 8 based on the risks generated from our model (represented as ROC AUC@5 and ROC AUC@8\textsuperscript{26}). In addition, using the R-based timeROC package with the rpy2 package, we evaluated the precision and recall of the models by generating AUC values indicating the extent of false positives and false negatives in the prediction\textsuperscript{g}. To visualize classification performance of the CNN model and to examine how they correlate with clinical features, we used t-SNE from the scikit-learn package in Python\textsuperscript{h}. Matplotlib\textsuperscript{i} was used for plotting all analyses.

**Multi-task Learning Analysis** We constructed contingency tables of the true and predicted values of drusen size and pigmentation abnormalities for both the fine-tuned ResNet152 and EfficientNetB3 models. Overall accuracy, sensitivity, specificity, and precision were calculated for each class. Drusen sensitivity, specificity, and precision were calculated in a one-vs-all method, where the identified value was considered positive and the other two values considered negative.

### 2.6 Analyzing Clinical Features

Two distinct datasets were used with both linear and non-linear models to predict the risk of patients developing late-AMD. The first set, labeled as clinical set A, contained age, smoking status, and for each eye, drusen size and presence of pigment abnormalities. These are the features used to calculate the 5-step simplified severity scale. Clinical set B contained age, smoking status, and for each eye: area of drusens within a central grid supplied to the grader, geographic atrophy (GA) within the central grid, subretinal GA atrophy, subretinal fibrosis, non-drusenoid pigment epithelial detachment, serous sensory retinal hemorrhaging, subretinal or subRPE hemorrhaging, RPE depigmentation, and increased pigmentation within the central grid. Drusen size and presence of pigmentation abnormalities in dataset A are calculated by binning features in dataset B to offer a more immediate and interpretable interpretation of the wellbeing of the patients’ eyes.

Features from either clinical set A or set B were extracted from the AREDS datasets. In a similar manner to the image features, the clinical features were either concatenated when analyzing multiple visits and the training sets were used to fit the same CoxPH model as described above. No PCA was used for clinical features. Survival analysis only included clinical features or "deep-features." No datasets contained both.

### 2.7 Hyper-parameter Tuning

Concordance index, ROC AUC at year 5, ROC AUC at year 8, precision-recall AUC at year 5, and precision-recall AUC at year 8 (as described in Evaluation Metrics) on the validation set were calculated for a wide range of learning rates on each combination of the model and dataset using 5-fold cross validation. The best learning rate for each combination was chosen as the learning rate which had the largest product of the mean of these five measures across the 5-fold cross validation in the validation set. All performance metrics shown are evaluated on the independent test sets during cross-validation.

### 3 Results and Discussion

#### 3.1 Longitudinal data improves risk prediction with clinical set A

With the limited features available in clinical set A, which are the features used in the 5-step simplifies severity scale, incorporating longitudinal data improves predictive performance (Table 1). Here, we see that the CoxPH with visits at years 0, 2, and 3 performs better than the CoxPH model using only data from year 3. We see a similar result with the MLP, where incorporating longitudinal data improved on the performance in comparison to using the single time point. Interestingly, the concatenated features performed better than the LSTM. As a whole, using longitudinal data with the clinical A dataset seems to be more informative than using just a single time point and the linear model performs equally as well as or better than the deep learning models.

\textsuperscript{g}https://cran.r-project.org/web/packages/timeROC/timeROC.pdf  
\textsuperscript{h}https://scikit-learn.org  
\textsuperscript{i}https://matplotlib.org/
### 3.2 Deep learning improves risk prediction with clinical set B

Clinical set B proves to be more informative than clinical set A in predicting a patient’s risk of developing late-AMD using all models (Table 2). In the more informative clinical set B, the deep learning models outperform the linear CoxPH model. Dataset B comprises many more features than dataset A, and these features are more “raw,” as in they are evaluated directly by the readers at the grading centers. It is possible that the deep learning models find more indicative interactions between the raw features than the transformation to dataset A does. The binning and dimensionality reduction for dataset A lost a large amount of useful information compared with the 22 features that are provided in dataset B. In this dataset, incorporating multiple timepoints only offers very slight improvements over using the single time point, indicating that the current time point is much more indicative of the patients’ risk of developing late-AMD than the previous time points.

### 3.3 Pretrained Features are not informative, but incorporating longitudinal data helps

ResNet152 and EfficientNetB3 pretrained features do not perform well (Table 3). The models that extract these features are tuned to classify dogs, birds, boats, and other types of natural image categories. The eye is a very delicate and intricate organ, where small sized abnormalities can cause a large difference. Therefore, the pretrained models are not capable of deciphering the sophisticated features that distinguish a healthy eye from a diseased eye. For both of these feature sets, the deep learning models outperform the CoxPH model. This is especially true in the case of the ResNet152 pretrained features because the PCA was only able to account for 22.4% of the variance in the feature set.

### 3.4 Fine-tuned features are more informative than the features they are trained on

The fine-tuned ResNet152 and EfficientNetB3 outperformed the clinical features and the pretrained features (Table 4). The models were trained on most of the features from dataset A yet outperformed dataset A in a survival setting. This
indicates that the deep learning models are able to extract more intricate features than drusen size and pigmentation, which can be more informative in predicting the risk of a patient developing late-AMD than drusen size and pigmentation. Figure 2 displays t-SNE plots of the feature vectors generated from the images used in the survival analysis (years 0, 2, and 3). Individual images are color coded by the amount of time between when that image was taken and when the patient reaches late-AMD. Dark red indicates that the patient is very soon to reach late-AMD, while white dots show that the patient has 8+ years until reaching late-AMD. The gray spots show the overall distribution of image features for both censored and uncensored patients. These t-SNE plots show that the deep feature extractors that are trained on drusen size and presence of pigmentation abnormalities are not only able to predict drusen size and presence of pigmentation abnormalities but are also able to extract features useful in predicting the risk of developing late-AMD without being explicitly trained to do so. These features are not only informative in predicting if a patient will develop late-AMD, but also in predicting if a patient will not develop late-AMD.

![Figure 2](image_url)

**Figure 2:** t-SNE plots of the fine-tuned deep features generated from the visits used in the survival analysis (years 0, 2, and 3). Colors indicate time to reach late-AMD. Dark red dots are soon to reach late-AMD. White dots will not reach for many years. Gray spots show the overall distribution of image features censored and uncensored patients.

### 3.5 LSTM improves short term predictions in fine-tuned features

Incorporating longitudinal data from the fine-tuned features does not aid in predicting all time risk of late-AMD. Only the most recent visit is necessary. However, incorporating longitudinal data and modeling the time dependencies of the data using the LSTM model improved on the more immediate prediction, shown by the large increase in PR AUC at year 5 for both the ResNet152 and EfficientNetB3 fine-tuned feature sets.

### 3.6 Multi-task Learning Performance

Both the EfficientNetB3 and ResNet152 were able to accurately classify drusen size and presence of pigmentation abnormalities. The EfficientNetB3 architecture more accurately predicted drusen size than the ResNet152 architecture. Both models had similar accuracies for predicting pigmentation abnormalities. As shown in Figure 3, both of the fine-tuned models are able to accurately cluster the different classifications. Additionally, in both models, we see a gradient from 0 to 2 in predicting drusen size. Class 2 drusen sizes are larger than class 1 and class 0 is the smallest. The models
are not explicitly told that this is the case, as each classification is treated as its own binary class. This shows that the fine-tuned CNN models are able to extract valid image features for evaluating fundus photographs.

![Figure 3: t-SNE plots of the feature vectors generated from the testing set used to evaluate the multi-task learning classifier models. Coloring indicates either the true (Reading Center Gradings) or predicted (Model Prediction) values for classifying drusen size and pigmentation abnormalities.](image)

**Table 5:** Results of the multi-task learning classifiers on drusen size (macro-average) and pigment abnormality.

<table>
<thead>
<tr>
<th></th>
<th>Drusen</th>
<th>Pigment Abnormality</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ResNet152</td>
<td>EfficientNetB3</td>
</tr>
<tr>
<td>Overall Accuracy</td>
<td>0.696</td>
<td>0.736</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.676</td>
<td>0.717</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.844</td>
<td>0.864</td>
</tr>
<tr>
<td>Precision</td>
<td>0.681</td>
<td>0.735</td>
</tr>
</tbody>
</table>

**3.7 Discussion**

In this study, a multi-task learning framework was used to predict drusen size and pigmentation abnormalities. By visualizing these features using t-SNE and calculating accuracy, sensitivity, and specificity for the model, our results show similar performance to state-of-the-art methods for these classification tasks. Interestingly, our features also correlate with the time-to-event prediction, where there was a clear separation for patients that have already reached late stage AMD to those patients who have not. The benefit of using our fine-tuned model, as compared to a pretrained model using Imagenet is shown in the significant increase in model performance based on evaluation metrics in the survival models. In addition, we show that EfficientNetB3, a newer developed model with fewer parameters, as compared to the ResNet152 model, showed higher performance. This may be due to the fact that the EfficientNet models were shown to better capture fine image details within an image. Further, we improved upon the standard CoxPH model that is commonly used in previous literature using deep learning methods. Last, while incorporating longitudinal data we do see improvements in short-term prediction performance and in clinical feature performance as compared to using a single time point data, but do not see improved long-term performance. It is interesting to note that when year 3 features performed better than years 0-3 features, the data from year 0 showed a decline in model performance (data not shown). This may indicate that more recent time points, for both clinical and image features, are more predictive than earlier time points in predicting risk of developing late-stage AMD.

While the biological changes that occur during AMD are well understood, unfortunately, there is no cure. This model can be used to aid clinicians in predicting how at risk patients are of developing late-AMD. Patients with low risk can have a treatment plan that will decrease patient costs and decrease the burden of patient care on the healthcare system.
In contrast, patients with high risk can receive a more aggressive treatment plan, at an earlier point in the disease to prolong vision as long as possible.

This study remains limited. Primarily, of the cohort analyzed in the survival analysis, only 350 patients reached late-AMD and results were not validated on an external dataset. Additionally, measures can be taken to improve the interpretability of the model, such as analyzing saliency maps to see which features are deemed most indicative in the multi-task learning models. To expand the model, the multi-task learning classifier could be extended by increasing the number of tasks learned. Adding in confidence scores would allow this model to aid grading centers by both reducing the time needed to analyze simple cases and assist in edge-case classification. Finally, although we explored a wide range of hyper-parameters and architecture makeups, we plan to conduct a more exhaustive analysis of MLP architectures and other hyper-parameter tunings.

4 Conclusion

This study shows that multi-task learning can be used to extract image features that are highly predictive of developing late-AMD. These extracted features are more predictive than the expert grader acquired feature, which are labor intensive and expensive to generate. This model can be used to aid clinicians in the stratification of patients with early-stage AMD, based on the risk of progressing to late-stage AMD. This would ease the exhaustive burden on the experts in the grading centers and greatly reduce cost. Additional future directions include integrating clinical features, such as smoking and age, and image features into the same deep learning model to try to improve model performance.

This model architecture is applicable to many other eye related diseases, including longitudinal prognosis of glaucoma. Additionally, the model could be extended far beyond fundus photographs to aid in longitudinal evaluation of non-eye related diseases such as cancers, COVID-19 and other diseases and illnesses.

Acknowledgements

The work was supported by the intramural program funds and contracts from the National Center for Biotechnology Information/National Library of Medicine/National Institutes of Health, the National Eye Institute/National Institutes of Health, Department of Health and Human Services (Contract HHS-N-260-2005-00007-C; ADB contract NO1-EY-5-0007; Grant No 4R00LM013001; NSF 1750326; NIH NIMH R01MH124740; NIH NIA RF1AG072449).

References

8. Yifan Peng, Shazia Dharssi, Qingyu Chen, Tiarann D. Keenan, Elvira Agrón, Wai T. Wong, Emily Y. Chew,


Simulating Screening for Risk of Childhood Diabetes: The Collaborative Open Outcomes tooL (COOL)

Mohamed Ghalwash, PhD1, Eileen Koski, M.Phil1, Riitta Veijola, MD, PhD2, Jorma Toppari, MD, PhD3, William Hagopian, MD, PhD4, Marian Rewers, MD, PhD5, Vibha Anand, PhD6

1 Center for Computational Health, IBM Research, NY, USA; 2 Department of Pediatrics, PEDEGO Research Unit, University of Oulu and Oulu University Hospital, Oulu, Finland; 3 Department of Pediatrics, Turku University Hospital, Turku, Finland; 4 Pacific Northwest Research Institute, Seattle, WA, USA; 5 Barbara Davis Center for Diabetes, University of Colorado, Denver, CO, USA; 6 Center for Computational Health, IBM Research, Cambridge, MA;

Abstract The Collaborative Open Outcomes tooL (COOL) is a novel, highly configurable application to simulate, evaluate and compare potential population-level screening schedules. Its first application is type 1 diabetes (T1D) screening, where known biomarkers for risk exist but clinical application lags behind. COOL was developed with the T1DI Study Group, in order to assess screening schedules for islet autoimmunity development based on existing datasets. This work shows clinical research utility, but the tool can be applied in other contexts. COOL helps the user define and evaluate a domain knowledge-driven screening schedule, which can be further refined with data-driven insights. COOL can also compare performance of alternative schedules using adjusted sensitivity, specificity, PPV and NPV metrics. Insights from COOL may support a variety of needs in disease screening and surveillance.

1 Introduction

The Collaborative Open Outcomes tooL (COOL) was designed to help researchers and clinicians evaluate different potential screening schedules simulating their impact on optimal case identification rates prior to symptom onset using retrospective datasets. The initial application has been in type 1 diabetes (T1D), however, the tool can be used for a variety of conditions in which changes in biomarkers may signal progression to disease.

T1D is a complex, heterogeneous, autoimmune disorder in which insulin-producing pancreatic beta cells are mistakenly destroyed by the body’s immune system. T1D has both genetic and familial components. Patients with T1D remain insulin dependent for life and are at high risk for serious long-term complications such as heart and kidney disease and diabetic retinopathy. For reasons that are not understood, T1D incidence rates have been rising dramatically. There is currently no cure or established prevention strategy for T1D and since newly diagnosed patients often present with diabetic ketoacidosis (DKA), a life-threatening condition with potentially long-term consequences, research on prevention and early detection is increasingly critical.

There are known biomarkers for T1D, however, progression to diabetes is heterogeneous, with generally a 1 to 5 year horizon from birth, which has complicated efforts to establish a practical screening paradigm. This pre-symptomatic period presents an opportunity for both improved prediction and prevention. The NIH-funded TEDDY study is currently investigating environmental determinants of T1D. Intervention trials of novel therapeutic agents focused on preventing or delaying (TrialNET and Innodia consortium), increase the importance of early identification of at-risk patients for trial recruitment. These efforts, combined with the known clinical risk of DKA at onset, means that early population-based screening could make important contributions to both research and practice.

The T1DI study group is a collaboration among IBM Research, JDRF and five of its academic partners: in the US (DAISY, DEW-IT, Sweden/DiPiS, Finland/DIPP) and Germany (BABYDIAB/BABYDIET). These studies have followed newborns and young children from the general population who are at genetic or familial risk for up to a period of 15 years for islet autoimmunity development or until diagnosis of clinical diabetes. The goal of the T1DI Study is to use advanced machine learning and statistical methods to derive novel insights into T1D disease processes and to develop risk stratification methods from the integrated T1DI cohort of over 24,000 subjects (with >2.5 million visits), approximately 3% of whom eventually received a T1D diagnosis. The T1DI cohort represents the largest data set for any natural history T1D study to date and the study aims to inform both research and clinical practice.
Figure 1: Screening for multiple biomarkers at age 5. COOL has 10 main panels - P1-P10. P1 defines the evaluation cohort. P2 defines the screening schedules. P3 allows the user to evaluate and explore the screening results. P4 shows the screening performance. P5-P7 show high level information about the cohort. P8 shows the number of subjects by category based on screening results. P9-P10 show further insights that could guide the user to refine the screening.

Cohorts, such as those included in this study, are expensive to maintain, and population-based screening also faces logistical obstacles in the case of a disease with heterogeneous onset patterns spanning many years. To complicate matters, geographical and socio-demographic factors also affect outcomes. The challenge for the practitioner and the researcher is to understand the nature of their own data enough to devise an effective screening paradigm.

Given the number and potential variety of variables involved, this poses a computational challenge for most clinicians and clinical researchers, and we knew of no tools available for this purpose. We developed COOL to help address this need, first applying it to our real-world case study of risk screening for T1D using the T1DI cohort. We have worked closely with our clinical collaborators to assure that the tool offers them the flexibility they need to iteratively ask “what-if” questions to determine how varying screening frequencies and ages can affect their ability to identify potential cases in a timely manner, and tuned to their specific population. This tool is currently being used actively in the T1DI Study in our efforts to define potential screening schedules for T1D.

2 Method

COOL is designed to allow users to simulate screening schedules for a given population to maximize identification of at risk patients based on known biomarkers. Four questions from clinical investigators in the T1DI collaboration will illustrate the utility of this highly configurable tool to gain new insights, and iteratively generate new hypotheses.

2.1 Research Question 1: When to screen subjects for biomarkers?

Clinical investigators start with a hypothesis such as: early detection of multiple biomarkers is likely to identify subjects with high risk of disease. Once validated, they must determine when screening should occur. COOL provides an intuitive interface, shown in Figure that helps users define, evaluate, and interactively explore the results of proposed screening schedules. Panels 1-3 allow the user to define a proposed schedule and panels 4-8 show the results for the available population. Panels 9-10 show additional insights that can guide the user to further refinements. These panels are described in more detail in the context of our results below.
COOL can also help evaluate and compare performance metrics for proposed screening schedules by providing functionality to address common challenges with retrospective cohorts. For example, a clinician may want to evaluate if screening for multiple biomarkers at the age of 5 years is sensitive enough to identify at-risk children. The question is domain knowledge driven, but the user may require special analytic techniques to utilize available data since not every subject may have been tested for known biomarkers at age 5. COOL can examine a subject’s data for a biomarker sample within a 6-month window before and after the target age. If found, biomarker test results classify the subject as positive or negative. If a subject has no biomarker sample in the given window, the test is deemed missing. Subjects’ biomarkers are thus categorized as: positive, negative, or no test. To evaluate screening performance given these categories, we compute cumulative sensitivity ($\text{sen}$), dynamic specificity ($\text{spc}$), positive predictive value ($\text{ppv}$), and negative predictive value ($\text{npv}$). These metrics are computed as:

$$\text{sen} = \frac{tp}{tp + fn + np}, \quad \text{spc} = \frac{tn}{tn + fp + nn}, \quad \text{ppv} = \frac{tp}{tp + fp}, \quad \text{npv} = \frac{tn}{tn + fn}$$

where $tp$ is the number of diagnosed subjects diagnosed who tested positive for the biomarker. $tn$ is the number of diagnosis-free subjects who tested negative. $fp$ and $fn$ are the numbers of false positives and false negatives, respectively. $np$ is the number of diagnosed subjects with no test and $nn$ is the number of diagnosis-free subjects with no test. Note that the denominator of sensitivity includes diagnosed subjects who had no test, i.e. $np$. This penalizes screening sensitivity if many subjects were diagnosed before the proposed screening age, but enables the metric to capture data on as many subjects as possible.

Right-censored data from subjects who are lost to follow up pose challenges in evaluating screening schedules because their outcomes are often unknown. However, excluding these subjects introduces bias in risk estimation for disease onset. To account for censoring, we use inverse probability weighting or IPCW mechanism, which assigns weights to subject data based on probability of censoring. It assigns larger weights to subjects with known diagnosis to offset those for whom it is unknown by the study end. For example, if the probability of censoring after 10 years of followup since birth is 0.2, it means that for any subject diagnosed by age 10, there are on average 4 subjects censored before age 10 and one subject followed through age 10. IPCW assigns a weight of 5 to the followed subject to offset the 4 censored subjects. Thus when computing $tp$, $tn$, $fp$, $fn$, $np$, and $nn$ metrics, each subject $i$ is assigned weight $w_i$. This built-in feature enables users to get unbiased results despite incomplete outcome data.

2.2 Research Question 2: How to improve screening performance?

The objective of screening is to improve early identification of at-risk subjects, but which screening strategy will be the most effective: a specific biomarker? any biomarker? multiple biomarkers? Can we modify the testing age? or add an additional testing age to obtain more sensitive (and specific) screening?

COOL allows the user to configure screening strategies by age and compare across different ages. In addition, the tool is flexible enough that a user can configure age-specific screening tests and chain them, such as: screening for any biomarkers at age 2 and for a specific biomarker at age 5. Figure 2 shows how COOL computes the results of a chained screening strategy. The result is positive if the subject tests positive at the first or second age or both. Subjects with no tests are categorized as missing. All other subjects (both negative or one negative and one missing) are categorized as negative. The advantage of this approach is that it can be applied recursively to chain any number of different screening tests at different ages.

2.3 Research Question 3: Should we confirm positive screening tests?

It is common practice to confirm positive (biomarker) test to reduce the impact of false positives although the confirmation strategy may vary by test. Subjects who test positive typically require subsequent tests. For example, when a subject tests positive for multiple biomarkers in the screening sample, should the confirmatory test focus on the initial positive biomarker(s) or all biomarkers? This question was raised to consider different confirmatory strategies to use in practice. COOL provides 5 strategies for confirmation, shown in Figure 3a. These policies are illustrated in Figure 3b for the screen for any biomarker strategy. The table shows sample status of four biomarkers in screening (blue columns) and confirmatory (green columns) samples. Subject S1 initially tested positive for two biomarkers,
Figure 2: Chain screening tests.

Figure 3: (a) COOL provides five confirmation options. *No confirmation*: no confirmation required for positive screens. *Confirm for any *IAb*: positive screen confirmed if any confirmatory sample biomarker is positive. *Confirm for at least one initial positive *IAb*: positive screen confirmed only if at least one positive screening biomarker is positive in confirmatory sample. *Confirm for at least two positive *IAb*: positive screen is confirmed only if at least two positive screening biomarkers are positive in confirmatory sample. *Confirm for all initial positive *IAb*: the positive screen is confirmed only if all positive screening biomarkers are positive in confirmatory sample. (b) Blue columns show the status of four biomarkers, B1-B4, for subjects S1-S6. Green columns show biomarker status in the confirmatory sample. 1=positive(+), 0=negative(-). Screening results (+/-) are shown for each subject and confirmation strategy.

B1 and B4. If *no confirmation* strategy is used, then screening results would be positive for S1. The *Confirm for any biomarker* strategy confirms positive screen if any biomarker is positive in the confirmatory sample regardless of the biomarker status in the initial sample. Under this policy, S4 confirmed positive although only B2 was positive in the first sample and only B1 was positive in the second (confirmatory) sample. The *Confirm for at least one initial positive biomarker* strategy confirms positive only if at least one of the positive biomarkers in the first sample is confirmed in the confirmatory sample. For example, S2 would screen positive under this strategy because it had two positive biomarkers B1 and B4 in the first sample, and one of them, B1, was confirmed positive. Similarly, the *Confirm for at least two initial positive biomarkers* strategy confirms positive screen only if at least two biomarkers are confirmed positive. Finally, the *Confirm for all initial positive biomarkers* policy confirms positive screen only if all positive biomarkers in the first sample are confirmed. The user can create more confirmation policies that suit the problem domain (it is beyond the scope of this paper to explain how the user can configure the tool).

2.4 Research Question 4: Is the screening universal or population-specific?

A fundamental attribute of screening guidelines is whether or not they can be applied universally. For example, Finland has the highest incidence of T1D, while The United States has the second highest incidence in North America. Is
the screening population-specific or universal? To answer this question, the same screening schedule can be applied to different cohorts with different characteristics, comparing performance in different populations. COOL provides a built-in cohort construction panel. The user can choose the evaluation cohorts in panel (1) in Figure 1 and compare screening performance on each chosen cohort. Whenever a screening schedule is evaluated, its results are appended to the table in panel (4) so users can compare performance across cohorts.

3 Results on Type 1 Diabetes

The following results are based on the T1DI cohort where islet autoantibodies were measured during 15-year follow-up or until T1D diagnosis.

3.1 Answer to Research Question 1: When to screen subjects for multiple biomarkers?

We use panel (1) to select the desired cohort and age, starting with age 5 years (Figure 4a), and to evaluate screening on the demo cohort for T1D risk by age 15. In panel (2) (Figure 4a) the user defines 3 screening parameters: target age, test and confirmation. Various screening strategies are implemented in COOL, such as multiple biomarkers, any biomarker, or specific combinations of biomarkers. These screening tests are shown in Figure 4b. For example, screening for any GADA tests the presence of Glutamic Acid Decarboxylase Autoantibodies in the blood sample. Screening for Only IAA+IA2 tests for the presence of Insulin Autoantibodies (IAA) and Insulinoma Antigen-2 Autoantibody (IA-2A) in the absence of other antibodies (GADA). The screening tests are not hard coded but are configurable by the tool making it easy to add additional tests. In the current use case, we screen for multiple biomarkers. Let us start with the hypothesis of no confirmation to reduce the number of tests and based on the results we can refine these parameters.

![Figure 4: (a) Panels 1-2: parameters to define a screening schedule. (b) Different screening strategies based on three biomarkers IAA, GADA, and IA2.](image)

After the user defines the screening schedule they can evaluate and explore the results. The tool immediately computes the performance of the screening schedule to assess how likely subjects are to develop disease by the designated age. The table in panel (4) shows the results of the chosen screening as in Figure 5a. Sensitivity, specificity, PPV, and NPV metrics are computed for evaluation of the schedule.

Showing the accuracy of the screening test is valuable to estimate how likely the screened-positive subjects are to develop the disease. If the results are not satisfactory the user can alter the proposed schedule to obtain improved results. The tool displays helpful insights about the results to guide the user about possible changes to the screening schedule or strategy. Panel (8) (Figure 5b) shows the number of subjects screened positive (red group, n=227), screened negative (green group, n=2779), and not tested (blue group, n=2994). The blue group did not have biomarker samples at the screening age. In each subgroup, the panel shows the number of cases and controls between parentheses. This information highlights the fact that there are many subjects with missing tests. So, if the user wants to capture these subjects they may want to screen at a different age, but what age needs to be determined?
3.2 Answer to Research Question 2: How to improve screening performance?

The previous section insights suggest that adding another test at an early age may improve screening performance. The user refines the screening schedule to two tests, at age 2 and 5 in panel (1), as shown in Figure 7a. The table in Figure 7b appends the results of the current screening strategy to the previous one for easy user comparison. From here, they can see that screening twice may be helpful - at ages 2 and 5 increases sensitivity from 0.35 to 0.51 and specificity from 0.51 to 0.56. The user can repeatedly chain different screening tests until improved results are obtained.

The results show that screening for multiple biomarkers twice at ages 2 and 5 is likely to identify subjects who will develop type 1 diabetes by age 15, but, can we improve performance further without an additional screening age, for example by exploring different testing strategies? COOL allows the user to explore different strategies as shown in Figure 4b. Figure 8 compares two screening policies at same ages but with different tests, screening for any biomarker versus screening for multiple biomarkers at ages 2 and 5. Screening for any biomarker has 13 points higher sensitivity than screening for multiple biomarkers, yet retaining similar positive predictive value. A user in a policy making capacity can perhaps use these results to begin to optimize the most effective screening schedule and tests for the intended population. In addition, the tool is flexible enough to allow the user to experiment. For example, screen for any biomarkers at age 2 and then screen for a specific biomarker at age 5 (Figure 9).
3.3 Answer to Research Question 3: Should we confirm positive screen?

As shown in Figure 10, various confirmation strategies are evaluated for screening for any IAb. The results indicate that confirming for any IAb has comparable sensitivity to no confirmation but a slightly higher PPV.

3.4 Answer to Research Question 4: Is the screening universal or population-specific?

The domain-knowledge driven screening schedule can be applied to each cohort, and results for each can be compared to assess whether the chosen screening schedule is applicable only to the selected cohort or universally, i.e. to all sub-cohorts in the available dataset. Furthermore, the user can evaluate them based on various domain attributes, e.g. male vs. female, high vs. low-risk HLA etc.

4 Discussion

We developed a pragmatic tool, (COOL), to evaluate efficacy of simulated screening strategies for chronic conditions. We applied COOL to screening for development of autoimmunity in progression to T1D onset. In this scenario, past research has provided important insights from discovery of immune biomarkers in the pre-symptomatic phase of the disease. These insights have yet to be translated into practice, specifically for screening children at risk for diabetes. However, the lack of screening in practice can be attributed to multiple factors, including unknowns of cost effectiveness and psychological readiness of those being screened, but the consequences are dire if not addressed expediently. This report also stresses that the natural history of the disease is still not fully understood, particularly varying ages and rates of progression, and the influence of ethnicity and environment. A tool such as COOL can help users explore the underlying data from large natural history studies to unlock potential explanations.

Using the example of T1D in 15 years of follow up and by leveraging existing data on autoimmunity development from the T1DI study group, we have shown how COOL can simulate population-level screening. We have shown four practical applications of this tool for evaluating screening strategies. A user can craft one or more competing “knowledge-driven” screening strategies. Then, the user can refine these strategies in a data-driven way by exploring the impact of underlying population distributions. These explorations can lead to further insights to improve risk screening performance, e.g. to maximize “at-risk catchment”. Screening strategies can be further improved by “chaining” different screening tests by screening age, testing sequence, or from what may be otherwise known in the domain.
Different strategies can be compared and a host of other alternative strategies devised for further exploration. Based on previously mentioned comparisons, a user can fine tune strategies for intended risk screening, e.g. universal or population specific. Within the framework of strategy development, a user can refine performance for varying time horizons (e.g T1D onset in 5, 10, or 15 years) as the tool accounts for right censored data in an unbiased way. Lastly, the built-in workbench capabilities of the tool retains alternatively evaluated strategies, which a user can recall later and further refine, for example when more data become available. The tool can also do a brute force search (evaluate all possible combinations configured) to find the optimal strategy.

In many ways the T1DI cohort is a unique dataset for understanding screening requirements, and for development of enabling tools. Its large sample size, variations in follow-up schedules among constituent studies and recruitment of newborns from both the general population and children and adolescents with family history have helped with experimentation on the domain problem as well as the tool design. The data lends itself to understanding variations in population-level screening policies to yield early results. This may be important because large scale screening studies in the US and EU are currently ongoing. The TEDDY study, which has rigorously followed at-risk children from young ages for development of autoimmunity and T1D is currently ongoing. There is no doubt TEDDY will shed more light on currently unknown disease processes and consequently inform screening policies but it will not reach the 15 year follow up mark until 2024. In the interim, as we have shown, evaluation of competing screening policies with existing T1D cohorts may help in the T1D domain. All stakeholders such as clinicians, policymakers, patient families may be able to pose pertinent questions, generate new hypothesis to test in-silico with COOL and get rapid answers. These advances will ultimately pave the way for faster adoption of screening in practice.

COOL differs from other tools in many respects. It is a general purpose and highly configurable evaluation tool, with built-in cohort selection and exploration capabilities that gives the user guidance based on underlying data and questions. Similar tools have generally focused on one question at a time, for example, risk prediction in a given time horizon, or evaluating drug design metrics or a web-based application for health-screening and have rarely sought to add functionality such as “chaining” alternate or competing strategies for risk evaluation and, to the best of our knowledge, none has sought to retain and compare all user-defined strategies in one tool.

However, as with all work, this tool also has some limitations. We do not address evaluation of repeated screening (or monitoring) strategies for a single patient (personalized), though we aim to investigate it in the future using state of

<table>
<thead>
<tr>
<th>Screening Test</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>PPV</th>
<th>NPV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Any IAb @ 2y No confirmation</td>
<td>0.64</td>
<td>0.54</td>
<td>0.43</td>
<td>0.99</td>
</tr>
<tr>
<td>Any IAb @ 5y No confirmation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Any IAb @ 2y Confirm for any IAb</td>
<td>0.64</td>
<td>0.55</td>
<td>0.48</td>
<td>0.98</td>
</tr>
<tr>
<td>Any IAb @ 5y Confirm for any IAb</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Any IAb @ 2y Confirm for at least one positive IAb</td>
<td>0.42</td>
<td>0.55</td>
<td>0.48</td>
<td>0.96</td>
</tr>
<tr>
<td>Any IAb @ 5y Confirm for at least one positive IAb</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Any IAb @ 2y Confirm for at least two positive IAb</td>
<td>0.49</td>
<td>0.55</td>
<td>0.48</td>
<td>0.97</td>
</tr>
<tr>
<td>Any IAb @ 5y Confirm for at least two positive IAb</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Any IAb @ 2y Confirm for all positive IAb</td>
<td>0.49</td>
<td>0.55</td>
<td>0.48</td>
<td>0.97</td>
</tr>
<tr>
<td>Any IAb @ 5y Confirm for all positive IAb</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 10: Evaluation of different confirmation strategies.
the art machine-learning (e.g. reinforcement learning) approaches and incorporate it into the tool. In this work, we also specifically aimed at risk screening policies in the pre-symptomatic period of a disease. As currently there is no readily available data on cost or yield, the cost effectiveness of screening (to prevent disease or cost of testing) was not evaluated. However, we believe that when these data become available, COOL can be enhanced to attach cost to the results of a screening test for evaluating cost-effectiveness.

5 Conclusion

In conclusion, the present analysis shows that Collaborative Open Outcomes tool (COOL) can unlock invaluable datasets, particularly from natural history studies of chronic conditions, to evaluate the efficacy and propel adoption of screening programs and potentially even inform current clinical practices. One example may be in a similar autoimmune condition - inflammatory bowel disease, where there are known biomarkers but no general population screening in practice. The tool’s capabilities can be applied to many forms of real world evidence such as data from electronic health records from hospitals, clinics and physician practices, as well as disease registries. A tool such as COOL can greatly facilitate exploration and evaluation of potential screening paradigms by clinical staff, using available data, which can have important implications for both research and clinical practice.

6 Acknowledgements


References


On the explainability of hospitalization prediction on a large COVID-19 patient dataset

Ivan Girardi, PhD1†, Panagiotis Vagenas, MS1†, Dario Arcos-Díaz, PhD2, Lydia Bessaï, MS2, Alexander Büsser, MS3, Ludovico Furlan, MD4, Raffaello Furlan, MD5, Mauro Gatti, PhD6, Andrea Giovannini, PhD1, Ellen Hoeven, MS2, Chiara Marchiori, PhD1
1IBM Research Europe, 2IBM GBS Germany, 3IBM GBS Switzerland
4Fondazione IRCCS Ca’ Granda, Ospedale Maggiore Policlinico, Milano, Italy
5Department of Biomedical Sciences, Humanitas University and IRCCS - Humanitas Research Hospital, Milano, Italy, 6IBM GBS Italy

Abstract
We develop various AI models to predict hospitalization on a large (over 110k) cohort of COVID-19 positive-tested US patients, sourced from March 2020 to February 2021. Models range from Random Forest to Neural Network (NN) and Time Convolutional NN, where combination of the data modalities (tabular and time dependent) are performed at different stages (early vs. model fusion). Despite high data unbalance, the models reach average precision 0.96-0.98 (0.75-0.85), recall 0.96-0.98 (0.74-0.85), and $F_1$-score 0.97-0.98 (0.79-0.83) on the non-hospitalized (or hospitalized) class. Performances do not significantly drop even when selected lists of features are removed to study model adaptability to different scenarios. However, a systematic study of the SHAP feature importance values for the developed models in the different scenarios shows a large variability across models and use cases. This calls for even more complete studies on several explainability methods before their adoption in high-stakes scenarios.

Introduction
As recently published1 during the COVID-19 pandemic, several AI-based models have been developed, targeting a large variety of application areas, such as identification of disease clusters, monitoring of cases, prediction of the future outbreaks, mortality risk, diagnosis of COVID-19, disease management, or vaccine development2. A thorough review found that almost all prediction models were trained on small or low-quality datasets, poorly reported and at a very high risk of bias or overfitting3. Due to the sense of urgency and the lax regulatory landscape, some of these models were even used without clinical validation, raising legitimate concerns about patient safety and model clinical validity4.

A way to build more robust AI prediction models is to adhere to the PROBAST framework, which provides a structured approach to assess potential risk of bias and applicability to the intended population and setting5. Transparent reporting can be achieved by following the TRIPOD checklist6. While the validation of the model "as-is" on external datasets is perceived as a proof of the generalization power of the model by some, this may be very difficult due to intrinsic differences in available populations, collected data, protocols and guidelines. Instead, a definitive important step forward would be to validate the models in real clinical settings by SMEs while in parallel improving their interpretability and correcting their bias with the help of explainability methods.

In this paper, we develop hospitalization prediction models on a large cohort of COVID-19 case data collected in the US from March 2020 to February 2021 and comprising more than 110k positive-tested patients (113941 upon data extraction, 110996 after further preprocessing). Knowing which patients are at risk for hospitalization for COVID-19-related symptoms or complications can help physicians decide not only how to best manage a patient’s care from the time of testing (e.g. remote monitoring, more frequent encounters), but also how to allocate resources. All the models reach relatively high performance despite the high class unbalance and the noise, typical of large real world datasets. We find however that despite the similar performance, different models lead to very different explanation results. While some hospitalization prediction models have been published7–10, they are generally built on smaller cohorts and do not systematically explore explainability methods across different AI models and feature subsets.

†Equal contributions
Methods

In this section we describe in details the various components of our approach: (i) Cohort definition; (ii) Data extraction: sourcing and preparation of data with partition of patient history in time intervals to enable increased flexibility during feature extraction; (iii) Feature extraction: accurate preparation and selection of features with special attention to data leakage prevention, missing data imputation, and proper feature encoding for capturing the temporal dimension of the data; (iv) Model development, where we experiment with time dependent and independent models (early and model fusion of temporal and Boolean modalities), (v) Internal testing and adaptability of the model for the use in several settings, by reducing the number of features used to train the model or varying the time interval between predictor assessment and outcome determination, and (vi) Model explainability.

Cohort definition

Our analysis was performed on the IBM Explorys Electronic Health Record (EHR) Database[1], which contains real-world clinical, operational, and financial data, spanning various healthcare aspects, from ambulatory to inpatient to specialty care. Explorys data are sourced from integrated delivery networks (IDN), clinically integrated network providers (CIN), and care collaborative groups (CC) and are constantly updated with incoming records. Data are curated, standardized, and normalized by using medical international coding standards, classification systems, ontologies, and lab unit measures such as ICD-10, SNOMED, LOINC, and RxNorm. Furthermore, data are searchable through a de-identified database. Explorys data have been used by the medical community for several studies, such as recently for building a predictive model for chronic kidney disease[2].

We used data collected from March 2020 to February 2021. To draw evidence of COVID-19 infection we used diagnoses and observations. We defined two evidence classes for COVID-19 diagnoses, confirmed and suspected, and examined the diagnosis fields ICD-10 code, the SNOMED concept, and a short diagnosis description snippet available in our source EHR data. As confirmed we defined the diagnoses explicitly registered with the ICD-10 code “U07.1”, which was introduced by the WHO in April 2020 for use “when COVID-19 has been confirmed by laboratory testing irrespective of severity of clinical signs or symptoms”. Diagnoses with the ICD-10 code “U07.2” (for use “when COVID-19 is diagnosed clinically or epidemiological but laboratory testing is inconclusive or not available”) were classified as suspected. Where the above COVID-19-specific codes were absent, we searched for a set of specific COVID-19-associated text patterns in the diagnosis description as well as in SNOMED concept name (SNOMED search was performed in a hierarchy-aware manner, as described later. Pattern matches were classified as suspected, unless detected by a set of exclusion patterns we defined, in which case the diagnoses were altogether disregarded from further analysis. To extract observation-based COVID-19 evidence, we used LOINC “94500-6”, which is the recommended code when (i) the gene or region being tested is not specified and a qualitative result is being reported (e.g., Detected/Not detected); or (ii) a single qualitative overall result based on a combination of individual test results is being reported, and also by far the most frequently occurring COVID-19 diagnostic test LOINC code in the data. Observations registered with LOINC code “94500-6” and positive result were defined as positive-test[3].

As cohort of interest we defined the one consisting of patients who had both a COVID-19 diagnosis, whether suspected or confirmed, and a positive-test observation within a time frame from 4 weeks prior to their earliest COVID-19 diagnosis up to 4 weeks following it. This cohort finally comprised 113941 patients. A statistic summary can be seen on Table[1]. Restricting the cohort only to patients with confirmed COVID-19 diagnosis and positive-test observation in the same time frame as above did not lead to significantly different results. We excluded patients who had a COVID-19 diagnosis but no positive-test observation, in order to reduce noise and focus on cases documented more completely in our EHR data. The absence of an explicit positive-test, is indeed ambiguous and can be explained either by the patient having been diagnosed by other means, e.g. chest CT, or as missing information, e.g. case not properly documented, test done in other institutions or unavailable (selection bias was not studied in detail, but the demographic distribution when including such “diagnosis-only” patients was similar to that of our defined cohort). A schema of the cohort selection is given in Figure[1].

*Testing availability was very limited until March-April 2020.
Figure 1: Left: cohort selection process. Right: Time partitioning of patient’s history referenced to the COVID-19 diagnosis and collected features over time intervals (Blue). Selection of features over the period of interest for an exemplary patient hospitalized 6 days after the diagnosis (Colored regions).

Data extraction

For each patient belonging to the cohort, we extracted demographics (age, gender), medical conditions (chief complaints, symptoms, comorbidities), vital signs and laboratory measurements, based on literature findings and SME input. With the exception of the demographics, all pieces of information were extracted with their corresponding timestamp. To capture the temporal dimension of the data, but also to reduce complexity, each patient journey was partitioned into various time intervals using the patient’s earliest COVID-19 diagnosis as reference point. Each variable was then extracted in an aggregated manner for each interval. The partitioning is depicted in Figure 1 and was designed to allow a greater focus from the diagnosis time onward.

To maximize recall, the same approach based on SNOMED and text pattern matching used to extract the COVID-19 diagnosis was used to extract symptoms, comorbidities and other medical conditions. In the case of SNOMED search, we performed a hierarchical pattern matching including also matched ancestors and using exclusion patterns to reduce false positives. We assigned a Boolean values to each medical condition in each time interval. Measurements and test results were extracted from the observation data, using corresponding LOINC codes and were normalized into a single unit. For each observation we extracted one, in most cases continuous, numeric variable per time interval, representing the patient’s most recently observed value in that interval, as well as the minimum, maximum, average. The patterns and codes used for data extraction were compiled in cooperation with SMEs.

Table 1: Basic statistics of our COVID-19 cohort in terms of patient gender, age, and location.

<table>
<thead>
<tr>
<th></th>
<th>Number of patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>63101 (55.38%)</td>
</tr>
<tr>
<td>Male</td>
<td>50827 (44.61%)</td>
</tr>
<tr>
<td>Unknown</td>
<td>13 (0.01%)</td>
</tr>
<tr>
<td>Age</td>
<td></td>
</tr>
<tr>
<td>Below 10</td>
<td>2612 (2.29%)</td>
</tr>
<tr>
<td>10 to 20</td>
<td>7952 (6.89%)</td>
</tr>
<tr>
<td>20 to 30</td>
<td>17206 (15.10%)</td>
</tr>
<tr>
<td>30 to 40</td>
<td>17401 (15.27%)</td>
</tr>
<tr>
<td>40 to 50</td>
<td>16744 (14.70%)</td>
</tr>
<tr>
<td>50 to 60</td>
<td>19019 (16.69%)</td>
</tr>
<tr>
<td>60 to 70</td>
<td>16417 (14.41%)</td>
</tr>
<tr>
<td>70 to 80</td>
<td>10116 (8.88%)</td>
</tr>
<tr>
<td>Above 80</td>
<td>6222 (5.46%)</td>
</tr>
<tr>
<td>Unknown</td>
<td>352 (0.31%)</td>
</tr>
<tr>
<td>State</td>
<td></td>
</tr>
<tr>
<td>Ohio</td>
<td>50948 (44.71%)</td>
</tr>
<tr>
<td>Louisiana</td>
<td>50638 (44.44%)</td>
</tr>
<tr>
<td>Illinois</td>
<td>5213 (4.58%)</td>
</tr>
<tr>
<td>Others (each &lt; 3.00%)</td>
<td>7142 (6.27%)</td>
</tr>
<tr>
<td>COVID-19 cohort</td>
<td>113941 (100.00%)</td>
</tr>
</tbody>
</table>

Figure 2: Number of COVID-19 cases by number of weeks since January 1, 2020.
In order to determine if a patient had been hospitalized in a given time interval, we examined the patient interactions with the healthcare providers, as captured in the patient’s encounter data. A patient was considered hospitalized in a time interval $T$, if they had an encounter with (i) encounter type inpatient, hospital emergency room visit or hospital encounter, (ii) admission date within $T$, AND (iii) duration above 24 hours. To exclude cases potentially hospitalized due to reasons other than COVID-19, patients with a hospitalization within the 4 weeks prior to the COVID-19 diagnoses were removed. After this cleanup step, patients with a hospitalization in the 4 weeks following the COVID-19 diagnosis were considered hospitalized due to COVID-19; we will refer to these patients simply as “hospitalized”. Interestingly, 79% of the hospitalized patients were admitted within 4 days following the COVID-19 diagnosis. The prediction target of our models was based on these COVID-19 hospitalization labels. It must be noted that some patients may be hospitalized in other institutions not part of the Explorys network, producing noise in form of false negatives. A possible reason could be bed unavailability due to sudden spikes in demand. The temporal distribution of hospitalized and non-hospitalized cases is illustrated in Figure 2.

Data preparation and feature extraction

Possible bias in the data was removed by discarding all features related to ethnicity, race, medical insurance type, location and any other available information not related to the medical conditions of the patient. Used feature types were demographics (age, gender), clinical presentation (symptoms, vitals, other medical conditions, comorbidities) and lab measurements. These categorical and continuous variables were associated to specific time intervals as previously described. We removed patients in case age information was unavailable or if all other features were empty. Missing values were replaced using data imputation methods, and features were encoded (e.g., one hot encoding for the age groups), after randomly splitting the data as 70% train set and 30% test set. More precisely, missing values for Boolean features were imputed with false, while continuous features were imputed with median substitutions learned on the train set. After feature imputation, we performed data normalization using average and standard deviation computed on the train set. For comorbidities and chronic conditions, mentions over all the available time intervals were aggregated using disjunction (“OR”) into a single Boolean feature (condition is present). For the other feature types, in order to avoid data leakage we selected only the time intervals from 14 days prior to the diagnosis up to the hospitalization (period of interest) with an upper limit of 28 days after COVID-19 diagnosis. Labs and vitals were kept with the corresponding time information, while all other Boolean medical conditions including symptoms were aggregated over the period of interest. We are aware that from a medical perspective the knowledge of whether a symptom was present for one or more consecutive days would be important to build the model, unfortunately our data did not consistently contain this type of information. Finally, for some conditions (i.e., stroke) we used the time intervals to distinguish between conditions belonging to the medical history of the patient (“past”) or to the period of interest. After the preprocessing steps described above the cohort comprised a total number of patients [110996], non-hospitalized [97082, %87] and hospitalized [13914, %13]. Since most patients were hospitalized within 4 days from the COVID-19 diagnosis date and given the sparsity of the features over time, we explored two approaches that combine the two data modalities (time series and tabular) at different stages in the model. In the first approach, the features from the two modalities were concatenated (early fusion) in an ordered feature vector $X = [X_1, X_2]$ where $X \in \mathbb{R}^{n \times k}$ with $n$ and $k$ the total numbers of patients and features, respectively. Data modality $X_1 \in \mathbb{R}^{n \times h}$ contains tabular data such as gender, age groups, comorbidities and other Boolean medical conditions, for a total of $h$ time-independent features. Data modality $X_2 \in \mathbb{R}^{n \times m \times t}$ contains $m$ features collected over $t$ time intervals, such as lab values and vitals. Concatenation of the modalities was performed after reshaping the second modality to a two dimensional matrix. In the second approach, the modalities $X_1$ and $X_2$ are fed into two different architectures, and their hidden representations are subsequently merged in the model (model fusion). Upon the described preprocessing, the total number of features $k = m \times t + h$ is $k = 1573$ with $h = 77$, $m = 88$ and $t = 17$.

Statistics

Statistical analysis was performed to compare the hospitalized ($H_1$) and non-hospitalized ($H_0$) populations and compute the significance on the rejection of the null hypothesis that there is no relationship between predictor and outcome. Demographic information, symptoms, comorbidities and other medical conditions are presented (as %) in Table 3, while median and IQR values for lab measurements and vitals are presented in Table 4. Comparison of the $H_1$ and
year old patients. A higher percentage of patients with symptoms were hospitalized (78.3%) of the 28.3% of the patients for which symptoms are actually known. Several of the conditions more commonly associated with older age: eighty year old patients are 4 times higher in nicotine dependence, chronic kidney disease). Our analysis shows increasing hospitalization admissions associated with progression (e.g. pneumonia, hypoxia, hypoxemia) or comorbidities and risk factors (e.g. hypertension, diabetes, nicotine dependence, chronic kidney disease). 

When comparing the percentages of symptoms and comorbidities reported in Table 2 to existing statistics, it should be noted that these numbers were calculated on the preprocessed dataset upon removal of all features collected after hospitalization. Large variability is generally observed across publications, likely due to differences in the cohorts, the way data were collected and the time the patient met the caregiver in their disease progression journey (e.g. at triage vs hospitalization). In the laboratory-confirmed US cohort described in Table 2, cough is reported in about 50% and fever in 43% of the 28.3% of the patients for which symptoms are actually known. Several of the conditions more commonly found in H1 in Table 2, are in agreement with literature findings, either as conditions correlated to severe disease progression (e.g. pneumonia, hypoxia, hypoxemia) or comorbidities and risk factors (e.g. hypertension, diabetes, nicotine dependence, chronic kidney disease). Our analysis shows increasing hospitalization admissions associated with older age: eighty year old patients are 4 times higher in H1 than H0, and 2 times higher for seventy and sixty year old patients.

Regarding the lab values reported in Table 3, it should be noted that these values are recorded before hospital admission.

Table 2: Demographic information, pre-existing comorbidities and acute conditions in % on COVID-19 positive patients. Comparison of non-hospitalized (H0) and hospitalized (H1) groups is performed with a \( \chi^2 \) contingency analysis with post-hoc correction (Benjamini–Hochberg procedure) is shown in the table with two asterisks when p-values < 0.001. Gender is one for males and zero for females.

H0 groups was performed with a \( \chi^2 \) contingency analysis for the Boolean feature and Mann–Whitney U test for the numerical ones. In our analysis we used post-hoc correction with the Benjamini–Hochberg procedure. Two asterisks were assigned when p-values are < 0.001.
where diagnoses of acute conditions are not available and the results of some very specific tests not accessible, we adaptability to different use cases, we removed selected lists of features. To emulate general practitioner settings (GP), we have defined a custom loss and weighted random sampling of the training data to take into account the high class unbalance.

To reduce overfitting. Comparing to the architecture trained by the Auto AI model with three dense layers followed by batch Stochastic Gradient Descend with momentum, learning rate was varied with a step scheduler and dropout used the number of epochs and let the library to find the best performing model. NN and T-CNN were optimized using mini iterations. The number of samples required to split an internal node, the minimum number of samples required to be at a leaf node parameters such as number of trees, number of features to consider for the best split, maximum tree depth, minimum split of the dataset into train, validation, and test sets was performed several times varying the seed.

During training we performed 3-fold cross validation and selected the best performing (best $F_1$-score) model on the validation set. Splitting of the dataset into train, validation, and test sets was performed several times varying the seed for the splitting. Precision, recall and $F_1$-score were collected by computing average and standard deviation over all iterations. RF and ET models were trained with a randomized parameter optimization to search for the optimal hyper-parameters such as number of trees, number of features to consider for the best split, maximum tree depth, minimum number of samples required to split an internal node, the minimum number of samples required to be at a leaf node either using the whole sample or bootstrapping. For the AutoKeras models we have varied the number of iterations and the number of epochs and let the library to find the best performing model. NN and T-CNN were optimized using mini batch Stochastic Gradient Descend with momentum, learning rate was varied with a step scheduler and dropout used to reduce overfitting. Comparing to the architecture trained by the Auto AI model with three dense layers followed by ReLu activation function and to emulate the same behavior of the NN (T-CNN) used in the training.

Prediction of hospitalization up to 28 days after COVID-19 diagnosis was modelled as a binary classification ($H_1$; hospitalized, $H_0$; not hospitalized). For early fusion approach we trained the following models: Random Forest classifier (RF), Extra Trees classifier (ET), fully connected neural network (NN) and Auto AI architectures (AutoKeras).

For the model fusion approach we trained an architecture combining NN for the time-independent modalities and temporal convolutional neural network (T-CNN) for the time-dependent one. The two models were combined together with additional NN layers. The T-CNN is a convolutional neural network consisting of 1-D temporal convolutions. During training we performed 3-fold cross validation and selected the best performing (best $F_1$-score) model on the validation set. Splitting of the dataset into train, validation, and test sets was performed several times varying the seed for the splitting. Precision, recall and $F_1$-score were collected by computing average and standard deviation over all iterations. RF and ET models were trained with a randomized parameter optimization to search for the optimal hyper-parameters such as number of trees, number of features to consider for the best split, maximum tree depth, minimum number of samples required to split an internal node, the minimum number of samples required to be at a leaf node either using the whole sample or bootstrapping. For the AutoKeras models we have varied the number of iterations and the number of epochs and let the library to find the best performing model. NN and T-CNN were optimized using mini batch Stochastic Gradient Descend with momentum, learning rate was varied with a step scheduler and dropout used to reduce overfitting. Comparing to the architecture trained by the Auto AI model with three dense layers followed by ReLu activation function and to emulate the same behavior of the NN (T-CNN) used in the training.

First, we built the models using all available features to have baselines (all features use case). Then, to test model adaptability to different use cases, we removed selected lists of features. To emulate general practitioner settings (GP), where diagnoses of acute conditions are not available and the results of some very specific tests not accessible, we

<table>
<thead>
<tr>
<th>Feature</th>
<th>Count $H_0$ (97082)</th>
<th>Count $H_1$ (13914)</th>
<th>Feature</th>
<th>Count $H_0$ (97082)</th>
<th>Count $H_1$ (13914)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HR (bpm/min)</td>
<td>49366</td>
<td>85.00 [75.00-96.00]</td>
<td>Glucose (mg/dL)</td>
<td>7813</td>
<td>114.00 [97.00-158.00]</td>
</tr>
<tr>
<td>SBP (mmHg)</td>
<td>46500</td>
<td>130.00 [118.00-141.00]</td>
<td>Neutrophils (10$^3$/µL)</td>
<td>4163</td>
<td>3.62 [2.55-5.28]</td>
</tr>
<tr>
<td>DBP (mmHg)</td>
<td>46494</td>
<td>78.00 [69.00-85.00]</td>
<td>CRP (mg/dL)</td>
<td>3075</td>
<td>5.30 [3.16-12.60]</td>
</tr>
<tr>
<td>T (°C)</td>
<td>50027</td>
<td>36.90 [36.70-37.20]</td>
<td>Ferritin (ng/mL)</td>
<td>2528</td>
<td>386.00 [160.00-917.00]</td>
</tr>
<tr>
<td>BMI (Kg/m$^2$)</td>
<td>49873</td>
<td>29.12 [27.70-34.45]</td>
<td>LDL (mg/dL)</td>
<td>2227</td>
<td>318.00 [220.00-487.75]</td>
</tr>
<tr>
<td>RR (breath/min)</td>
<td>39226</td>
<td>18.00 [16.00-19.00]</td>
<td>CK (UI/L)</td>
<td>2236</td>
<td>113.00 [61.00-211.00]</td>
</tr>
<tr>
<td>Hb (g/dL)</td>
<td>12417</td>
<td>13.20 [11.90-14.40]</td>
<td>D-dimer (µg/mL)</td>
<td>11053</td>
<td>111.50 [11.40-14.40]</td>
</tr>
<tr>
<td>혈액단위</td>
<td>12369</td>
<td>0.95 [0.80-1.23]</td>
<td>Creactive (mg/dL)</td>
<td>10898</td>
<td>1.05 [0.80-1.46]</td>
</tr>
<tr>
<td>PLT (10$^9$/µL)</td>
<td>11658</td>
<td>21.00 [18.50-26.00]</td>
<td>eTSTR (mg/dL)</td>
<td>2019</td>
<td>0.03 [0.01-0.07]</td>
</tr>
<tr>
<td>WBC (10$^9$/µL)</td>
<td>11656</td>
<td>5.63 [4.56-7.81]</td>
<td>ePTT (s)</td>
<td>1161</td>
<td>30.30 [27.40-34.80]</td>
</tr>
<tr>
<td>K (mmol/L)</td>
<td>11831</td>
<td>4.10 [3.70-4.30]</td>
<td>Eosinophils (10$^3$/µL)</td>
<td>2012</td>
<td>0.08 [0.04-0.14]</td>
</tr>
<tr>
<td>BUN (mg/dL)</td>
<td>11682</td>
<td>15.00 [11.00-21.00]</td>
<td>Basophils (10$^3$/µL)</td>
<td>1605</td>
<td>0.03 [0.03-0.05]</td>
</tr>
<tr>
<td>ALB (g/dL)</td>
<td>10794</td>
<td>3.80 [3.60-4.20]</td>
<td>PH</td>
<td>413</td>
<td>7.38 [7.31-7.44]</td>
</tr>
<tr>
<td>AST (U/L)</td>
<td>10416</td>
<td>30.00 [21.00-45.00]</td>
<td>CO2 (mmHg)</td>
<td>422</td>
<td>38.00 [32.00-46.00]</td>
</tr>
<tr>
<td>Monocytes (10$^3$/µL)</td>
<td>10149</td>
<td>0.50 [0.40-0.70]</td>
<td>HCO3 (mmHg)</td>
<td>413</td>
<td>7.38 [7.31-7.44]</td>
</tr>
<tr>
<td>Bilirubin (mg/dL)</td>
<td>10023</td>
<td>0.50 [0.30-0.70]</td>
<td>NonHb (g/dL)</td>
<td>3951</td>
<td>0.50 [0.40-0.70]</td>
</tr>
<tr>
<td>Lymphocytes (%)</td>
<td>10201</td>
<td>20.90 [13.30-30.30]</td>
<td>WBC (10$^9$/µL)</td>
<td>308</td>
<td>6.00 [4.00-11.00]</td>
</tr>
<tr>
<td>ALT (U/L)</td>
<td>9935</td>
<td>25.00 [17.00-38.00]</td>
<td>PTT (s)</td>
<td>375</td>
<td>13.70 [12.80-15.40]</td>
</tr>
<tr>
<td>Lactate (mg/dL)</td>
<td>9769</td>
<td>1.20 [0.80-1.70]</td>
<td>Bilirubin</td>
<td>199</td>
<td>1.00 [1.00-2.00]</td>
</tr>
<tr>
<td>SpO2 (%)</td>
<td>13798</td>
<td>98.00 [96.00-99.00]</td>
<td>Lactate (mmol/L)</td>
<td>50</td>
<td>1.00 [0.72-1.60]</td>
</tr>
<tr>
<td>Na (mmol/L)</td>
<td>5944</td>
<td>138.00 [115.00-140.00]</td>
<td>mean BD (mg/dL)</td>
<td>69</td>
<td>90.00 [79.00-102.00]</td>
</tr>
<tr>
<td>HR pulse asymmetry (%)</td>
<td>29584</td>
<td>98.00 [97.00-99.00]</td>
<td></td>
<td>5742</td>
<td>96.00 [94.00-98.00]</td>
</tr>
</tbody>
</table>

Table 3: Median and IQR of the time dependent features reported in the EHR close to the COVID-19 diagnosis. Comparison of non-hospitalized ($H_0$) and hospitalized ($H_1$) groups is performed with Mann–Whitney U test with post-hoc correction (Benjamini–Hochberg procedure) is shown in the table with two asterisks when $p$-values < 0.001.
removed the following: pneumonia, hypoxia, hypoxemia, myocarditis, all sepsis related conditions, acute respiratory distress syndrome, heart disease, pneumothorax, acute renal failure syndrome, thromboembolic disorder, acute hepatic failure, pulmonary embolism, renal failure syndrome, embolism, acute disease of cardiovascular system and lab tests such as D-dimer and hsTnT. This GP use case was designed under SMEs guidance. Finally, to test the model predictive power over time, we removed all available features in the day prior to the hospitalization for hospitalized patients (one day before use case). This drastically reduces the availability of lab values and vitals. Average precision, recall, $F_1$-score are given in Table 4 for the all feature use case, the GP use case and the one day before use case, as first, second and third values in the series, respectively. NN and T-CNN architectures were frozen across all use cases. In all described use cases, all analysed models achieved relatively high classification performance despite the unbalanced dataset. There is a minimal difference amongst the time dependent T-CNN model performance and the NN and the Auto Keras models for all features and the GP use cases. Drop of $F_1$-score when using a smaller number of features was limited to few percentages for all the models, at the exception of the T-CNN in one day before use case. The observed similar performances of the models may be due to the presence of several predictive features in the dataset, most likely, more than the dropped ones.

<table>
<thead>
<tr>
<th>Model</th>
<th>$P(H_0)$</th>
<th>$R(H_0)$</th>
<th>$F(H_0)$</th>
<th>$P(H_1)$</th>
<th>$R(H_1)$</th>
<th>$F(H_1)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>RF</td>
<td>0.97/0.97/0.96</td>
<td>0.98/0.97/0.98</td>
<td>0.98/0.97/0.97</td>
<td>0.83/0.79/0.84</td>
<td>0.82/0.81/0.74</td>
<td>0.83/0.80/0.79</td>
</tr>
<tr>
<td>ET</td>
<td>0.97/0.96/0.96</td>
<td>0.98/0.97/0.98</td>
<td>0.97/0.96/0.97</td>
<td>0.81/0.77/0.83</td>
<td>0.80/0.74/0.75</td>
<td>0.80/0.75/0.79</td>
</tr>
<tr>
<td>NN</td>
<td>0.98/0.97/0.97</td>
<td>0.96/0.95/0.95</td>
<td>0.97/0.96/0.96</td>
<td>0.74/0.72/0.72</td>
<td>0.85/0.82/0.79</td>
<td>0.79/0.77/0.75</td>
</tr>
<tr>
<td>T-CNN</td>
<td>0.98/0.98/0.97</td>
<td>0.96/0.95/0.94</td>
<td>0.97/0.96/0.96</td>
<td>0.75/0.71/0.67</td>
<td>0.83/0.84/0.83</td>
<td>0.79/0.77/0.74</td>
</tr>
<tr>
<td>Auto Keras</td>
<td>0.96/0.96/0.96</td>
<td>0.98/0.97/0.97</td>
<td>0.97/0.97/0.97</td>
<td>0.85/0.80/0.81</td>
<td>0.75/0.75/0.72</td>
<td>0.80/0.77/0.76</td>
</tr>
</tbody>
</table>

Table 4: Hospitalization prediction in terms of average precision (P), recall (R), $F_1$-score (F) within 28 days after the COVID-19 diagnosis by using (i) all available features prior to the hospitalization, (ii) carefully removing a selected list of features with guidance of SMEs, and (iii) removing all available features up to one day prior to the hospitalization.

To determine which features have the largest predictive power and assess model interpretability, we explored feature importance methods. Although several studies used MDI importance on datasets containing numerical features, we disregarded it, given its sensitivity to high-cardinality features. In addition, commonly used libraries compute MDI-based importance values on the training set and therefore the extracted important features are not useful to make predictions that generalize on the test set. Instead, we focused on the SHAP method and report feature importance values computed on the test set. Regarding the NN and T-CNN models, SHAP values computed with the gradient based approximation are very similar to the ones computed using different not-gradient-based approximations. Therefore we did not regularize the gradients during training as suggested by. The top 35 SHAP values computed on the RF and NN trained with all available features are given in Figure 3 (corresponding plots for ET and CNN have been omitted due to their similarity to the RF and the NN plots, respectively). Median of the absolute SHAP values and corresponding interquartile ranges (IQR) are reported with colored boxes, $1.5 \times$ IQR ranges with whiskers and mean of the absolute values with white circles. Although some features overlap (12/35, 34%), lab values and vitals are predominant in RF (60%) while a minority in the NN model (6%). This is even more accentuated after dropping the list of Boolean medical conditions in the GP use case (lab values 71% in RF vs 6% in NN, Figure 4). When removing all features up to one day prior to the hospitalization, the number of lab values available to the model drops consistently and in this case the overlap between important features for the two models increases up to 25/35 (71%).

For the Boolean conditions, a high level agreement is observed between the information of Table and the feature importance results shown in Figures. However, the statistical analysis is performed feature by feature and is fundamentally different from the feature importance model described. The latter is based on a predictive model which takes into account feature combination during the decision process and hence does not require features to be independent. In both cases, presence of relationship from the statistical analysis or importance from the SHAP method does not imply causality. Reason for hospitalization comes from combination of several factors.
Cerebrovascular disease (past)  Acute renal failure syndrome  Acute renal failure syndrome  Coronary heart disease

Chronic kidney disease  Lymphocytes (10³/ L)  Neutrophils (10³/ L)  Monocytes (10³/ L)  Platelets (10³/ L)  Renal failure syndrome  Lymphocytes (%)  Neutrophils (%)  Monocytes (%)  Platelets (%)  ARDS  Glucose (mg/dL)  Hemoglobin (g/dL)  Hyperlipidemia  Hypokalemia  ATPase (mmol/L)  Creatinine (mg/dL)  Cystatin C (mg/L)  Tachycardia  CRP (mg/dL)  WBC (10³/ L)  RR (breaths/min)  Heart rate (beats/min)  C-reactive protein (mg/dL)  Creatine (g/dL)  Albumin (g/dL)  AST (U/L)  ALT (U/L)  T ( °C)  SBP (mmHg)  DBP (mmHg)  Blood pressure (mmHg)  Pneumonia  Hypoxia  Hypertension  Hypoxemia  RR (breaths/min)  COPD  Smoking status  Gender  Hyperlipidemia  Heart disease  Nicotine dependence  Diabetes  Dyspnea  Asthma  Fatigue  Headache  Twentyes  Severe cases  Chest pain  Hypertension  Pregnancy  Cancer  Heart failure  Nausea  Chronic kidney disease  Anemia  Asthma  Asthma  Acute renal failure syndrome  Fatigue  Headache  Twentyes  Nausea  Chronic kidney disease  Stroke  Anemia  Asthma  Asthma  Cough  Hypertension  Cough  Diabetes

Figure 3: Median SHAP values for the RF (left panel) and NN (right panel) computed on the test sets (across all splits) and using all available features. Corresponding interquartile ranges (IQR) are reported with colored boxes, 1.5× IQR ranges with whiskers and mean values with white circles.

Figure 4: Same as Figure 3 but using a restricted number of available features. See text for further details.

Conclusions

We developed several AI models (RF, ET, NN, T-CNN) to predict hospitalization on a large cohort of COVID-19 patients (more than 110k) up to 28 days after the diagnosis, with competitive classification results on the test set. Despite the large data unbalance, using available features prior to the hospitalization, the models reach average 0.96-0.98 (0.75-0.85) precision, 0.96-0.98 (0.74-0.85) recall, and 0.97-0.98 (0.79-0.83) F1-score across all split iterations for H0 (H1). Similar results were obtained in more restrictive scenarios with lower number of available features, even when features ranked as top by importance models were removed from the baseline model (GP use case). We then systematically explored the SHAP feature importance method for all analysed models in all the use cases and observed an overall agreement with the results of the statistical analysis on the Boolean medical conditions. However, the observed
Figure 5: Same as Figure 3 but for RF and NN models after removing features.

large variability of the SHAP importance score across models calls for a careful usage of the importance results, their evaluation across different models and methods. When deriving feature importance, methods that can handle feature correlation like SHAP should be used or analysis of feature correlation carefully conducted. Undesired dependencies as the ones shown in this study need to be carefully evaluated before adoption in high-stakes decision.

Limitations

In this retrospective study we tried to adhere as much as possible to the PROBAST framework, in the constrained setting of this use case: hospitalization prediction requires a large dataset of hospitalized and non-hospitalized patients and therefore data sources of different provenience are needed (eg., GP office visits, emergency, hospital encounters). This means that if a patient left or suddenly came into the Explorys network for whatever reasons (e.g., hospitalized in other hospitals or regions), outcomes and data may be missing. This may lead also to large variability in laboratory values and measurements. Our models were not tested on external datasets or in clinical settings due time constraints.

References

22. Ross AS, Hughes MC, and Doshi-Velez F. Right for the right reasons: training differentiable models by constraining their explanations. Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17), 2017;2662–2670.
Weak Supervision for Affordable Modeling of Electrocardiogram Data

Mononito Goswami, B.Tech.1, Benedikt Boecking, M.Sc.1, Artur Dubrawski, Ph.D.1
1Auton Lab, School of Computer Science, Carnegie Mellon University
Pittsburgh, PA, USA

Abstract

Analysing electrocardiograms (ECGs) is an inexpensive and non-invasive, yet powerful way to diagnose heart disease. ECG studies using Machine Learning to automatically detect abnormal heartbeats so far depend on large, manually annotated datasets. While collecting vast amounts of unlabeled data can be straightforward, the point-by-point annotation of abnormal heartbeats is tedious and expensive. We explore the use of multiple weak supervision sources to learn diagnostic models of abnormal heartbeats via human designed heuristics, without using ground truth labels on individual data points. Our work is among the first to define weak supervision sources directly on time series data. Results show that with as few as six intuitive time series heuristics, we are able to infer high quality probabilistic label estimates for over 100,000 heartbeats with little human effort, and use the estimated labels to train competitive classifiers evaluated on held out test data.

Introduction

Automatic analysis of electrocardiograms (ECGs) promises substantial improvements in critical care. ECGs offer an inexpensive and non-invasive way to diagnose irregularities in heart functioning. Arrhythmias are abnormal heartbeats which alter both the morphology and frequency of ECG waves, and can be detected in an ECG exam. However, identifying and classifying arrhythmias manually is not only error-prone but also cumbersome. Clinicians may have to analyze each heartbeat in an ECG record, and in critical care settings, carefully analysing each heartbeat is nearly impossible. As a consequence, the medical machine learning (ML) community has worked extensively on computational models to automatically detect and characterize arrhythmias1,2.

Rajpurkar et al.3 demonstrated that modern ML models trained on a large and diverse corpus of patients can exceed the performance of certified cardiologists in detecting abnormal heartbeats. But their Convolutional Neural Network model was trained on a manually annotated dataset of more than 64,000 ECG records from over 29,000 patients. Clearly, research on automated arrhythmia detection has moved the burden of monitoring ECG in critical care to annotating and curating large databases on which ML models can be trained and validated. This currently prevailing process involves laborious manual data labeling that is a major bottleneck of supervised medical ML applications in practice. Popular ML techniques, in particular deep learning, require a large supply of reliably annotated training data, containing records from a diverse cohort of patients. According to Moody et al.4 and our own experience, raw medical data is abundant, but its thorough characterization can be involved and expensive. This reliance on labeled data forces researchers to often use static and older datasets, despite evolving patient populations, systematic improvements in understanding of diseases, and advances in medical equipment.

Recent developments in e.g. web-based tools to visualize and annotate ECG signals have not reduced the annotation time and effort significantly. For example, it took 4 doctors, almost 3 months to annotate 15,000 short ECG records using the LabelECG tool5. In general, gold standard expert annotations can be costly. Conservative estimates place the hourly cost of highly qualified labor for the related task of EEG annotation between $50 and $200 per hour6.

In this work, we explore the use of multiple cheaper albeit perhaps noisier supervision sources to learn an arrhythmia detector, without access to ground truth labels of individual samples. We follow the recently proposed data programming (DP)7 framework in which a factor graph is used to model user-designed heuristics to obtain a probabilistic label for each heartbeat instance. DP has gained attention from the medical imaging and general ML community and has been used for various tasks such as automated detection of seizures from electroencephalography8, intracranial hemorrhage detection with computed tomography, or automated triage of Extremity Radiograph Series9.

Our experiments with ECG data from the MIT-BIH Arrhythmia Database indicate that with as few as 6 heuristics, we are able to train an arrhythmia detection model with only a small amount of human effort. The resulting model
Figure 1: Data programming with time series heuristics can affordably train competitive end models for automated ECG adjudication. Instead of labeling each data point by hand (fully supervised setting), experts encode their domain knowledge using noisy labeling functions (LFs). A label model then learns the unobserved empirical accuracy of LFs and uses them to produce probabilistic data label estimates using weighted majority vote.

is competitively accurate when compared to a model trained on the same data with full supply of pointillistic ground truth annotations. It can also outperform another alternative model trained using active learning, a popular technique used to reduce data labeling efforts when they are expensive. We also show that domain heuristics can be automatically tuned to account for inter-patient variability and further boost reliability of the resulting models.

While many different types of arrhythmias exist, for illustration purposes we focus on identifying heartbeats showing Premature Ventricular Contractions (PVCs). Whereas isolated infrequent PVCs are usually benign, frequent PVCs with exceptionally wide QRS complexes† may be indicative of heart disease and eventually lead to sudden cardiac death\textsuperscript{10}. However, our approach is general and applicable to all classes of abnormal heartbeats.

Related Work

Automated Arrhythmia Detection Automatically detecting abnormal heartbeats is a widely studied problem. Most researchers in the past relied on manually labeled corpora such as the MIT-BIH Arrhythmia Database, the AHA Database for Evaluation of Ventricular Detectors, etc., to train and validate their models\textsuperscript{1, 2}. Rajpurkar et al.\textsuperscript{3} recently demonstrated that a deep Convolutional Neural Network (CNN) can even exceed the performance of experienced cardiologists. However, their model was trained on as many as 64,121 thirty-second ECG records from 29,163 patients, manually-labeled by a group of certified cardiographic technicians. Hence, to fuel advances in automated arrhythmia detection and, more generally, in ML-aided healthcare, there is a clear need to affordably label vast amounts of data.

Some recent studies have attempted to address the annotation bottleneck, albeit at a different context and scale. These studies have used semi-supervised or active learning to incrementally improve the accuracy of models without significant expert intervention. For instance, to overcome inter-patient variability without additional manual labeling of patient specific data, Zhai et al.\textsuperscript{11} iteratively updated the preliminary predictions of their trained CNN using a semi-supervised approach. Correspondingly, Wang et al.\textsuperscript{12} used active learning on newly acquired data to choose the most informative unlabeled data points and incorporate them in the training set. Sayantan et al.\textsuperscript{13} used active learning to improve their model’s classification results with the help of an expert. So far, the work which comes closest to addressing the problem of intelligently labeling vast quantities of ECG data is that of Pasolli et al.\textsuperscript{14}. Starting from a small sub-optimal training set, the authors proposed three active learning strategies to choose additional heartbeat instances to further train an Support Vector Machine (SVM) model. Their work demonstrated that models trained using active learning can achieve impressive performance while using few labeled samples. In this work, we also compare the performance of our weakly supervised method with active learning. As against Pasolli et al.\textsuperscript{14} who used a manually

†QRS complexes are generally the most prominent spike seen on a typical ECG. They are a combination of the Q wave, R wave and S wave, which occur in rapid succession, and represent an electrical impulse.
curated set of ECG features and trained an SVM using margin sampling, we train a Convolutional Neural Network (CNN) which can automatically learn rich feature representations using uncertainty sampling, another popular active learning strategy.

**Weak Supervision** Of late, some studies have explored the use of multiple noisy heuristics to programatically label data at scale. The recently proposed Data Programming framework\(^7\), where experts express their domain knowledge in terms of intuitive yet perhaps noisy labeling functions (LFs), is a prominent example. Recent studies have used DP for a wide range of clinical applications, ranging from detecting aortic valve malformations using cardiac MRI sequences\(^8\), seizures using EEG\(^9\) and brain hemorrhage using 3D head CT scans\(^9\). With the exception of Khattar et al.\(^10\), most prior work on DP has been on image\(^8,15\) or natural language\(^17\) modalities. Moreover, prior DP research either used weak annotations from lab technicians, students\(^8\), or heuristics built on auxiliary modalities (e.g., clinician notes, text reports\(^9\), patient videos\(^8\)), some of which only allow for coarse annotation of the entire time series rather than of the individual segments. On the contrary, we define heuristics directly on time series. This enables seamless labeling of entire time series or their segments using the same framework.

**Methodology**

In this section, we will describe how we use domain knowledge to define heuristics to detect PVC in ECG time series. These heuristics will noisily label subsets of data. We will model these noisy labels to obtain an estimate of the unknown true class label for each data point. We then use the estimated labels to train the final classifier, which will be evaluated on held out test data and compared to alternative models trained using ground-truth labels directly. Fig. 1 describes the full workflow we follow to train the end model \(f\).

**Domain Knowledge to Identify PVC** A Premature Ventricular Contraction is a fairly common event when the heartbeat is initiated by an impulse from an ectopic focus which may be located anywhere in the ventricles rather than the sinoatrial node. On the ECG, a PVC beat appears earlier than usual with an abnormally tall and wide QRS-complex, with its ST-T vector directed opposite to the QRS vector, Fig. 2(ii). These generic characteristics allowed one non-domain-expert user to define 6 heuristics in less than 30 minutes. The user was initially unfamiliar with clinical ECG interpretation and referred to an online textbook\(^18\) to develop heuristics. Expert clinicians are likely able to define heuristics more rapidly and thoroughly. The heuristics listed below are defined directly on time series. This is in contrast to prior work which uses weak annotations or heuristics defined on an auxiliary modality such as text or images.

**Heuristics:** i. R-wave appears earlier than usual, ii. R-wave is taller than usual, iii. R-wave is wider than usual, iv. QRS-vector is directed opposite to the ST-vector, v. QRS-complex is inverted, vi. Inverted R-wave is taller than usual.

**Modeling Labeling Functions over Patient Time Series** We will now describe our formal assumptions about the dataset and heuristics, and introduce the modeling procedure. Given an ECG dataset of \(p\) patients \(X = \{x^j\}_{j=1}^p\), where \(x^j \in \mathbb{R}^T\) are raw ECG vectors of length \(T\), we can segment each ECG \(x^j\) into \(B < T\) beats such that \(x^j = \{x^j_1, \ldots, x^j_B\}\). Each segment \(b \in \{1, \ldots, B\}\) has an unknown class label \(y_b \in \{-1, 1\}\), where \(y_b = 1\) represents a premature ventricular contraction (PVC). Our goal is to use domain knowledge to model the unknown \(y_b\), without having to annotate the instances individually, to then train an end classifier \(f(x^j_B) = y_b^j\) for automatic detection of PVC. We define \(m\) labeling functions (LFs) \(\{\lambda_h(x^j_i)\}_{h=1}^m\) directly on the time series. These LFs noisily label subsets of beats with \(\lambda_h(x^j_i) = \{-1, 0, 1\}\) corresponding to votes for negative, abstain, or positive. These LFs do not have to be perfect and may conflict on some samples, but must have accuracy better than random\(^9\). DP uses this voting behavior to infer true labels by learning the empirical accuracies, propensities and, optionally, dependencies of the LFs via a factor graph. We use a factor graph as introduced in Ratner et al.\(^7\) to model the \(m\) user defined labeling functions. For simplicity, we assume that the LFs are independent conditioned on the unobserved class label. Let \(Y^j = (y^j_1, \ldots, y^j_B) \in \{-1, 1\}^B\) be the concatenated vector of the unobserved class variable for the \(B\) beat segments of patient \(j\) and \(\Lambda^j = \{-1, 0, 1\}^{B \times m}\) be the LF output matrix where \(\Lambda^j_{ik} = \lambda_k(x^j_i)\) is the output of LF \(k\) on beat \(i\) of
patient \( j \). We define a factor for LF accuracy as

\[
\phi^{Acc}_{i,k}(\Lambda, Y) \triangleq 1\{\Lambda_{ik} = y_i\}
\]

We also define a factor of LF propensity as

\[
\phi^{Lab}_{i,k}(\Lambda, Y) \triangleq 1\{\Lambda_{ik} \neq 0\}
\]

Then, the label model for a patient \( j \) is defined as

\[
p_{\theta}(Y^j, \Lambda^j) \triangleq Z_{\theta}^{-1} \exp\left( \sum_{i=1}^{B} \sum_{k=1}^{m} \theta_k \phi^{Acc}_{i,k} (\Lambda^j_i, y^j_i) + \sum_{i=1}^{B} \sum_{k=1}^{m} \theta_k \phi^{Lab}_{i,k} (\Lambda^j_i, y^j_i) \right)
\]

where \( Z_{\theta} \) is a normalizing constant. We use Snorkel\(^{17}\) to learn \( \theta \) by minimizing the negative log marginal likelihood given the observed \( \Lambda^j \). Finally, as introduced in Ratner et al.\(^{7}\), the end classifier \( f \) is trained with a noise aware loss function that uses probabilistic labels \( \hat{Y}^j = p_{\theta}(Y^j|\Lambda^j) \).

**Figure 2:** Examples of a normal (i) and PVC (ii) heartbeat. Dotted green horizontal lines represent the ECG baselines detected during pre-processing, blue and red vertical lines mark the QRS-complexes and T-waves.

**From Domain Knowledge to an Automated Arrhythmia Detector** First, we minimally pre-processed the raw ECG signals by removing baseline wandering using a forward/backward, fourth-order high-pass Butterworth filter\(^{20}\).

To segment ECG \( (x^j) \) into individual beats \( (x^j_b) \), we followed a simple segmentation procedure, where we considered the time segment between two alternate QRS-complexes to be a heartbeat.

We had to determine the precise locations of the QRS-complexes and T-waves. As in most prior work, we used the approximate locations of the R-wave available for each ECG record in the database, along with Scipy’s peak finding algorithm\(^{21}\) to find the exact locations of the R and T waves. Further, we used the RANSAC algorithm\(^{22}\) to fit a robust linear regression line to each ECG record, to determine its baseline (horizontal green lines in Fig. 2). The baselines were used to accurately characterize the height and depth of the R and T-waves\(^{‡}\) (blue and red vertical lines in Fig. 2).

Next, we defined 6 simple LFs based on the domain knowledge to assign probabilistic labels (PVC, OTHER or ABSTAIN) to each beat. Fig. 4 provides example pseudocodes for two of the LFs that were defined. To express the loosely-defined domain knowledge we described previously as LFs, we have to automatically assign thresholds to them. For instance, one heuristic to identify a PVC beat is to check whether its “R-wave appears earlier than usual”. To turn this heuristic into a LF (LF\(_{Early \text{ R-wave}}\)), one has to determine the “usual” position of the R-wave. For this, we used the Minimum Covariance Determinant algorithm\(^{23}\) to find the covariance of the most-normal subset of the frequency histogram. We then set the threshold to the value 2 standard deviations away from the estimated mean in the direction of interest. For example, for a particular subject (Fig. 3) LF\(_{Early \text{ R-wave}}\) returns PVC for any beat with the R-wave appearing earlier than \( T_{Early \text{ R-wave}} = 238 \text{ ms} \) (vertical green line). To account for inter-patient variability, we automatically compute these subjective thresholds for each heuristic and every patient separately. Note that some of our heuristics did not require estimating any subject-specific parameters.

\(^{‡}\)The topological prominence measure returned by Scipy’s peak finding algorithm was imprecise.
To turn a loosely defined heuristic such as “R-wave appears earlier than usual” into an LF, we must characterize the “usual” location of the R-wave. To accomplish this, we we fit a robust Gaussian distribution to model the variance of R-wave locations, and assume any beat located two standard deviations earlier (solid green line at $T_{\text{Early R-wave}} = 292$ ms), than the estimated mean, to be a PVC beat. Since these attributes vary widely from patient-to-patient, we automatically compute these thresholds separately for each patient and heuristic.

The End-Model Classifier
With these heuristics, we use the label model in Eq. (1) to obtain probabilistic labels for heartbeats of all training patients in the MIT-BIH Arrhythmia Database. We use these probabilistic labels and the segmented beats to train a noise-aware ResNet classifier, in which we weigh each sample according to the maximum probability that it belongs to either class. Recent studies have shown that ResNet not only performs on par with most of the state-of-the-art time series classification models\(^{24}\), but also works well for automatic arrhythmia detection\(^1\).

Experiments and Results

Data
The Massachusetts Institute of Technology – Beth Israel Hospital (MIT-BIH) Arrhythmia Database\(^4\) is one of the most commonly used datasets to evaluate automated Arrhythmia detection models. It contains 48 half-hour excerpts of two-lead ECG recordings from 47 subjects. In most records, the first channel is the Modified Limb lead II (MLII), obtained by placing electrodes on the chest. We only used the first channel to detect PVC events, since the QRS-complex is more prominent in MLII. The second channel is usually lead V1, but may also be V2, V4 or V5 depending on the subject. We refer the interested reader to Moody et al.\(^4\) for more details on how the database was curated and originally annotated.

Experimental Setup
Our experimental setup follows the evaluation protocol of arrhythmia classification models stipulated by the American Association of Medical Instrumentation (AAMI) as described in\(^{25}\). The AAMI standard, however, does not specify which heartbeats or patients should be used for training classification models, and which for evaluating them\(^2\). Hence, we used the inter-patient heartbeat division protocol proposed by De Chazal et al.\(^{26}\) to partition the MIT-BIH Arrhythmia Database into subsets DS1 and DS2 to make model evaluation realistic. Furthermore, in the MIT-BIH Database, PVCs only account for $8\%$ of the 100,000 beats, thus to prevent issues stemming from high class imbalance, we randomly oversampled PVC beats in DS1 before using it to train the ResNet classifier. The
architecture of our ResNet models is the same as in Fawaz et al.\textsuperscript{24}. To tune the learning rate, batch size, and number of feature maps hyper-parameters, we split the training data into train and validation subsets in the 70/30 proportion. All the models were trained for 25 epochs. In the next subsection, we report the results of the ResNet models which had the best true positive rate (TPR) at low false positive rate (FPR) on the validation data.

We also compare the performance of our weakly supervised model with an active learning (AL) alternative. For AL, we used ResNet to iteratively identify data points for manual labeling using uncertainty sampling\textsuperscript{27}. The model initially had access to a randomly sampled balanced seed set of 100 labeled data points. In each AL iteration, we retrained ResNet using the training data extended with 100 newly labeled data points. We continued this process until the training set consisted of 4,000 points. AL hyper-parameters (the query size and size of the seed set) are similar to Pasolli and Melgani’s setup\textsuperscript{14}. Table 1 reports the performance of the final ResNet model trained on 4000 data points incrementally labeled using AL, averaged over 10 random initializations of the seed set.

<table>
<thead>
<tr>
<th>Model</th>
<th>TPR</th>
<th>TNR</th>
<th>PPV</th>
<th>FPR</th>
<th>Acc</th>
<th>FPR\textsubscript{50% TPR}</th>
<th>FNR\textsubscript{50% TNR}</th>
<th>TPR\textsubscript{1% FPR}</th>
<th>TNR\textsubscript{1% FNR}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fully sup.</td>
<td>0.884</td>
<td>0.970</td>
<td>0.664</td>
<td>0.030</td>
<td>96.25</td>
<td>0.005</td>
<td>0.028</td>
<td>0.793</td>
<td>0.266</td>
</tr>
<tr>
<td>Pr. labels</td>
<td>0.645</td>
<td>0.960</td>
<td>0.523</td>
<td>0.039</td>
<td>85.84</td>
<td>0.019</td>
<td>0.140</td>
<td>0.165</td>
<td>0.252</td>
</tr>
<tr>
<td>Active learn.</td>
<td>0.514</td>
<td>0.993</td>
<td>0.821</td>
<td>0.007</td>
<td>94.15</td>
<td>0.020</td>
<td>0.021</td>
<td>0.604</td>
<td>0.405</td>
</tr>
<tr>
<td>Weak sup.</td>
<td>0.892</td>
<td>0.965</td>
<td>0.629</td>
<td>0.036</td>
<td>97.25</td>
<td>0.004</td>
<td>0.013</td>
<td>0.707</td>
<td>0.466</td>
</tr>
</tbody>
</table>

Table 1: Results on held-out test set. Weakly supervised ResNet performs on par with the fully supervised model and outperforms ResNet trained using active learning. FPR\textsubscript{50% TPR} and FNR\textsubscript{50% TNR} represent the FPR and FNR at 50% TPR and TNR, respectively. Similarly, TPR\textsubscript{1% FPR} and TNR\textsubscript{1% FNR} represent the TPR and TNR at 1% FPR and FNR, respectively. The reported AL results are averaged over 10 independent initializations of the random seed set. All measures are computed with PVC as the positive class.

**Results** We trained ResNet models on DS1 as a training set, using either probabilistic labels or the full ground truth, and evaluated them on the held-out set DS2. The results, summarized in Tab. 1, reveal that the end classifier trained using weak supervision is competitive with the model trained on the full ground truth data. Moreover, our weakly supervised model also outperformed the ResNet trained using 4,000 data points obtained via active learning.

**Figure 5:** Active learning results. Our weakly supervised model either exceeds or matches the performance of its AL counterpart. The shaded red regions correspond to the 95% Wilson’s score intervals.

Let us review the key insights stemming from these results. First, the thresholds for the labeling functions that were automatically determined by our proposed auto-thresholding algorithm varied quite drastically across subjects. For instance, the threshold on the position of the R-wave, T\textsubscript{Early R-wave} had a mean of 230 ms and a standard deviation of 77.14 ms. This simple personalization of the LF parameters turned out to be the key to good generalization properties of the end-model; it failed to perform well when these parameters were fixed to reasonable global settings. The auto-thresholding algorithm is a practically important contribution of our work, as it allows our methods to scale across diverse cohorts of subjects while mitigating potentially excessive manual effort in tuning LFs to specific patients. However, unsurprisingly, even with auto-tuning, our LFs and the estimated probabilistic labels (denoted “Pr. labels” in Tab. 1) were not perfect. In fact, we observed high variability in the performance of Pr. labels across different subjects, when compared to ground truth. For example, while they had almost perfect sensitivity for Subject 228 (TPR = 0.994),
they performed extremely poorly for Subject 214 (TPR = 0). Overall, Pr. labels had low TPR and high TNR on their own on the training set and held-out test set, which is understandable given the prior class imbalance.

![Graphs](i) (ii) (iii)

Figure 6: Class Activation Maps for weakly [(i) - (iii)] and fully supervised [(iv) - (vi)] ResNet reveal that the models discriminate between PVC and OTHER beats primarily based on the morphology of the QRS-complex. The models appear to have learned to focus on similar regularities. Graphs (i) and (iv) represent an example OTHER beat, while the others show two examples of PVC beats.

Tab. 1 summarizes performance metrics on the held-out test set (True Positive Rate (TPR), True Negative Rate (TNR), Positive Predictive Value (PPV), False Positive Rate (FPR), and Accuracy (Acc)) measured at the 50-50 class likelihood threshold, chosen for consistency with prior literature on the PVC prediction task. The weakly supervised ResNet (Weak sup.) significantly improves sensitivity to the PVC class compared to just using the LF labels directly (Pr. labels) to directly predict the test data. This illustrates that WS ResNet is able to generalize effectively beyond the hypothesis learned by the noisy weak LFs. Our end model performs on par with the fully supervised ResNet (Fully sup.) trained on the same data but using all the available pointillistic labels in the MIT-BIH Database.

Tab. 1 also compares performance of the four models under consideration at operational settings of pragmatic interest in clinical practice, that is at very low error rates. We report the ability to confidently identify positive cases at FPR of 1%, and the ability to confidently identify negatives at FNR of 1%. We complement these results with the error rates observed at 50% probability of detection of both negative and positive cases. The results show very little operational utility potential from applying the inferred probabilistic labels directly. However, our weakly supervised ResNet model trained on those inferred labels is highly competitive to the equivalently structured ResNet trained on the abundant supply of manually annotated data. Weakly supervised ResNet appears particularly strong at identifying negative cases, while its positive recall performance is close to the ground-truth based equivalent.

Next, we compare performance of the weakly supervised ResNet versus ResNet trained using active learning (“Active learn.” in Tab. 1), and it looks better on all performance metrics barring TNR, PPV and FPR. Graphs in Fig. 5 show that in the range of up to 4,000 pointillistically labeled training data points, weakly supervised models either outperform or matches its active learning counterpart, but at a drastically lower requirement of human effort.

To closely examine what the weakly and fully supervised ResNet models are learning, we plotted Class Activation Maps\(^2^8\) of a normal and PVC beats in Fig. 6. It is evident that to discriminate between PVCs and other beats, our models are primarily paying attention to the QRS complexes in these examples. Moreover, it also appears from these plots that both models not only perform on par, but they also tend to focus on similar signatures of the ECG signals. This observation suggests at least some equivalence between the model trained on ground-truth annotation and the one trained on labels inferred from a small number of simple heuristics. These results reassure us that the more expensive process can be effectively replaced by the proposed framework of weak supervision that uses a few labeling functions.
based on high-level aspects of domain knowledge derived directly from the time series characteristics.

Discussion and Conclusion

We demonstrated that weak supervision with domain heuristics defined directly on time series provides a promising avenue for training medical ML models without the need for large, manually annotated datasets. To support this claim, we developed an arrhythmia detection model which performs on par with its fully-supervised counterpart, and does not need point-by-point data annotation. This weakly supervised model has been developed in a fraction of time that would be required to provide a fully labeled training set.

We only needed a handful of heuristics to infer probabilistic labels sufficient to yield a reliable end model. These simple heuristics reflected basic clinical intuition that can be gleaned from ECG diagnostics tutorials. We expect that engaging expert clinicians to harvest additional heuristics would allow further improvements. We stipulate that the proposed approach does not only save effort and time, but it also aligns the process of knowledge acquisition from domain experts better with human nature, than its tedious pointillistic data annotation alternative. Further, we show that domain heuristics can be automatically tuned to patient specific characteristics by defining parameter tuning rules. In our example, auto-tuning of ECG waveform interpretations accounts for inter-patient variability, while keeping manual labor at its minimum.

The ML community has devised several techniques to overcome the limitations of expensive pointillistic labeling such as intelligently choosing the most informative training samples to label, combining both labeled and unlabeled data and harnessing the power of crowds. While semi-supervised learning has been successfully applied to improve arrhythmia detection models without patient-specific data, these methods still rely on a significant proportion of labeled training data to start with. On the other hand, crowdsourcing has shown promise in generating ground truth for e.g. medical imaging, but prior research found several limitations such as the lack of trustworthiness, inability of non-expert workers to annotate fine-grained categories and ethical concerns around patient privacy. Active learning, however, has by far been the most commonly utilized technique in settings where annotating large quantities of data en-masse is prohibitively expensive.

Multiple avenues of future work include modelling dependencies between LFs to improve both the efficiency and accuracy of label models, and developing a library of time series primitives to streamline development of LFs for such data. We would also like to build interfaces to support interactive discovery of LFs and to rigorously validate resulting end models. Further, we intend to investigate hybrid approaches that will opportunistically combine weak supervision with pointillistic active learning, and conduct user studies with clinicians to better understand the challenges and opportunities for interactive harvesting of domain knowledge. We also aim to enable detection of other types of abnormalities that can be seen in ECG data, and apply our approach to other types of hemodynamic monitoring waveforms.

Time series data is prevalent in healthcare. However the costs of preparing such data for training and validation of new models, as well as for the maintenance of already developed models, prohibit the otherwise realizable benefits from widespread adoption of machine learning in clinical decision support. We believe that approaches similar to the one presented in this paper could help making a decisive push towards proliferating beneficial uses of machine learning in this important field of its application.

Acknowledgements

This work was partially supported by the Defense Advanced Research Projects Agency award FA8750-17-2-0130, and by the Space Technology Research Institutes grant from National Aeronautics and Space Administration’s Space Technology Research Grants Program.

References


Using Radiomics as Prior Knowledge for Thorax Disease Classification and Localization in Chest X-rays

Yan Han¹, Chongyan Chen¹, Liyan Tang¹, Mingquan Lin², Ajay Jaiswal¹, Song Wang¹, Ahmed Tewfik¹, George Shih³, Ying Ding¹,*, Yifan Peng²,*
¹The University of Texas at Austin, Austin, TX, USA; ² Population Health Sciences, Weill Cornell Medicine, New York, NY, USA; ³ Department of Radiology, Weill Cornell Medicine, New York, NY, USA;

Abstract

Chest X-ray becomes one of the most common medical diagnoses due to its noninvasiveness. The number of chest X-ray images has skyrocketed, but reading chest X-rays still have been manually performed by radiologists, which creates huge burnouts and delays. Traditionally, radiomics, as a subfield of radiology that can extract a large number of quantitative features from medical images, demonstrates its potential to facilitate medical imaging diagnosis before the deep learning era. In this paper, we develop an end-to-end framework, ChexRadiNet, that can utilize the radiomics features to improve the abnormality classification performance. Specifically, ChexRadiNet first applies a light-weight but efficient triplet-attention mechanism to classify the chest X-rays and highlight the abnormal regions. Then it uses the generated class activation map to extract radiomic features, which further guides our model to learn more robust image features. After a number of iterations and with the help of radiomic features, our framework can converge to more accurate image regions. We evaluate the ChexRadiNet framework using three public datasets: NIH ChestX-ray, CheXpert, and MIMIC-CXR. We find that ChexRadiNet outperforms the state-of-the-art on both disease detection (0.843 in AUC) and localization (0.679 in T(IoU) = 0.1). We make the code publicly available at https://github.com/bionpllab/lung_disease_detection_amia2021, with the hope that this method can facilitate the development of automatic systems with a higher-level understanding of the radiological world.

1 Introduction

The chest X-ray is one of the most common medical procedures for diagnosis, but the interpretation of chest x-ray images is subject to significant diagnosis variability for important clinical decisions. A radiologist reads about 20,000 images a year, roughly 50-100 per day, and the number is increasing. Each year, the US produces 600 billion images, and 31% of American radiologists have experienced at least one malpractice claim, often missed diagnoses¹. The shortage of radiologists and burnout of physicians creates an urgent demand for immediate solutions. Building automatic or semi-automatic approaches to medical imaging diagnosis becomes an unavoidable next step.

The recent development of artificial intelligence, especially deep learning, offers great potential to improve medical imaging diagnosis². It also sneaks into the radiology reading rooms to build a new paradigm for precision diagnosis³–⁵. Pioneering work on chest X-rays mainly focused on two problems: disease classification and localization. The recent release of large-scale datasets, such as NIH Chest X-ray⁴, CheXpert⁶, and MIMIC-CXR⁷, have enabled many studies using deep learning for automated chest X-ray diagnosis, such as thorax disease classification³,⁸–¹⁰ and localization⁸,¹¹,¹².

In practice, radiologists use pattern recognition on medical images to make a diagnostic decision¹³. The knowledge of radiologists can be captured by Radiomics, which has demonstrated the effectiveness of image-based biomarkers for cancer staging and prognostication. Formally, radiomics extracts quantitative data from medical images to represent tumor phenotypes, such as spatial heterogeneity of a tumor and spatial response variations. It plays an important role in precision medicine to support evidence-based clinical decision-making. For example, radiomics can generate the detailed quantification of tumor phenotype¹⁴ and acts as a radiographic imaging phenotype which is associated with tumor stage, metabolism, and gene or protein expression profiles¹⁵,¹⁶.

While radiomics offer the potential for more precise and accurate clinical predictions, it is surprising that radiomics has not been implemented in the layers of the neural networks, nor to the best of our knowledge in the deep learning

*Equal contributions.
Figure 1: Model overview. The model contains three major parts. Blue arrows represent the feedforward multi-label classification part. The below black arrows represents the mask generation and radiomic features extraction part. Red arrows means the radiomic features regularization and backward part.

workflow for X-ray analysis\textsuperscript{17,18}. To bridge this gap, in this paper, we propose ChexRadiNet, a new framework that incorporates domain-specific knowledge (radiomics) into deep learning algorithms as soft constraints, and then learns end-to-end to automatically detect thorax diseases and generate bounding boxes on chest X-rays. Compared with previous studies, our proposed model does not need pre-annotated bounding boxes for training and can achieve state-of-the-art performance for thorax disease localization. Therefore, it provides a way to introduce prior information about anticipated explanations, a technique that is widely used in the “Rationale model”\textsuperscript{19} (Section 2). For ensuring ChexRadiNet is robust and generalizable, three public benchmarking datasets were used for this purpose: NIH Chest X-ray\textsuperscript{4}, CheXpert\textsuperscript{6}, and MIMIC-CXR\textsuperscript{7}. We demonstrate that our model outperforms baseline methods for both thorax disease classification and localization (Section 3).

2 Method

Figure 1 shows our proposed ChexRadiNet, which consists of two branches. The first branch predicts whether the pathology is present or not in the image. The second branch localizes its regions using the radiomic features extracted from the first branch. ChexRadiNet utilizes a multi-task, closed-loop strategy to learn and use radiomic features as soft constraints. Formally, we are learning a two-part latent-variable model of the form \( E_{z \sim p(z|x)} p(y|x, z) \), where the latent \( z \) is a radiomic-based mask over the image \( x \) with the probability \( p(z|x) \). \( p(y|x, z) \) is a masked version of the classification framework. Therefore, we consider the training process as a weakly-supervised learning. In this section, we first illustrate the architecture of ChexRadiNet and then present the training process.

2.1 Model architecture

2.1.1 Branch I: Multi-label classification

In this branch, we label each image with a 14-dim vector \( y = [y_1, \ldots, y_k, \ldots, y_K] \), \( y_k \in \{0, 1\} \), \( K = 14 \) for each image. \( y_k \) indicates the presence with respect to the according pathology in the image while a zero vector represents the status of “Normal” (no pathology is found in the scope of any of 14 disease categories as listed).

We use the residual neural network (ResNet) architecture\textsuperscript{20}, given its dominant performance in ILSVRC competitions and the triplet attention mechanism (see Section 2.1.3). However, our framework can be applied to other CNNs. ResNet-18 and ResNet-50 are used in this paper. After removing the final classification layer and global pooling layer, an input image with shape \( h \times w \times c \) produces a feature tensor with shape \( h' \times w' \times c' \) where \( h, w, \) and \( c \) are the height, width, and number of channels of the input image, respectively while \( h' = h/32, w' = w/32, c' = 2048 \). The output of this network encodes the images into a set of abstracted feature maps. Then through an application of two convolutional layers (each followed by batch normalization and ReLU activation), the number of channels is modified
to $K$, where $K$ is the number of possible disease types. A perchannel probability for each disease class is then derived by a fully-connected layer with a sigmoid activation function; this is denoted $p(k|I)$, where the probability is that whether the image belongs to class $k$ and $I$ denotes the image. Since we intend to build $K$ binary classifiers, we will exemplify just one class $k$. Note that $k$th binary classifiers will use the $k$th-channel features to do prediction. Since all images have their labels, the loss function for class $k$ can be expressed as minimizing the binary-cross entropy as $L_k = -y_k \log p(k|I) - (1-y_k) \log(1-p(k|I))$, where $y_k$ is the ground truth label of the $k$ class. To enable end-to-end training across all classes, we sum up the class-wise losses to define the total loss as $L_I = \sum_k L_k$.

2.1.2 Branch II: Mask generation

In this branch, we generate bounding boxes (B-Box, or masks) based on the classification result of Branch I to get the most indicative areas using the class activation mappings (CAMs). The heatmap produced from the model indicates the approximate spatial location of one particular thoracic disease class each time. Due to the simplicity of intensity distributions in these resulting heatmaps, applying an ad-hoc thresholding-based B-Box generation method for this task is found to be sufficient. Followed by the work of Wang et al., the intensities in heatmaps are first normalized to $[0, 255]$ and then thresholded by $\{60, 180\}$ individually. Finally, B-Boxes are generated to cover the isolated regions in the resulting binary maps.

**Radiomic features extraction.** With the generated B-Boxes and original images, we extracted radiomic features to regularize the model. Quantitative radiomics can be categorized into the following subgroups:

- **First-order statistics features** describe the distribution of individual pixel values without concerns for spatial relationships. They are histogram-based properties using mean, median, maximum, and minimum values of the pixel intensities on the image, as well as their asymmetry, flatness, uniformity, and entropy.

- **Shape features** describe the shape of the region of interest (ROI) and its geometric properties (e.g., volume, maximum diameter along with different orthogonal directions, maximum surface, tumor compactness, and sphericity).

- **A Gray Level Co-occurrence Matrix (GLCM) features** describe the second-order joint probability function of an image region constrained by the mask. The matrix $P(i, j|\delta, \theta)$ represents the number of times the combination of levels $i$ and $j$ occurs in two pixels in the image, that are separated by a distance of $\delta$ pixels along angle $\theta$.

- **A Gray Level Size Zone (GLSZM) features** quantify gray level zones in an image. A gray level zone is defined as the number of connected pixels that share the same gray level intensity.

- **A Gray Level Run Length Matrix (GLRLM) features** quantify gray level runs, which are defined as the length in number of pixels, of consecutive pixels that have the same gray level value.

- **A Neighboring Gray Tone Difference Matrix (NGTDM) features** quantify the difference between a gray value and the average gray value of its neighbors within distance $\delta$. The sum of absolute differences for gray level $i$ is stored in the matrix.

- **A Gray Level Dependence Matrix (GLDM) features** quantify gray level dependencies in an image. A gray level dependency is defined as the number of connected pixels within distance $\delta$ that are dependent on the center pixel.

All above features can be extracted either directly from the images or after applying different filters or transforms (e.g., wavelet transform). In our design, we utilize the Pyradiomics tool to extract radiomic features (https://pyradiomics.readthedocs.io/).

Finally, we use the pairwise distance between radiomic features and image features as regularization. Therefore, the adjustable loss function is $L_{II} = L_I + \|I_F - R_F\|_p$, where $I_F$ and $R_F$ are the image features and radiomic features, respectively, and $\| \cdot \|_p$ denotes the norm and $p$ represents the norm degree, e.g., $p = 1$ and $p = 2$ represent the Taxicab norm and Euclidean norm, respectively. In this paper, we set $p$ to 2. Please note that although the original shapes of $I_F$ and $R_F$ are not equal, we easily adapted one-layer MLP to project them into the same dimension space.
2.1.3 Triplet Attention

To boost the quality of masks, we integrate the triplet-attention mechanism\(^ {22}\). Triplet Attention mechanism requires few learnable parameters and could capture important features by taking cross-dimension interaction into account\(^ {22}\). In other words, it includes three sub-branches to respectively capture the dependency between spatial dimensions Height (\(H\)), Width (\(W\)), and the Channel (\(C\)) dimension. For the first branch, in measuring the interactions between dimension \(H\) and dimension \(C\), it first performs a Z-pool operation by concatenating the result of average pooling and max pooling across dimension \(W\). This operation can be summarized as \(\chi_1 = z\text{-pool}(\chi') = [\text{MaxPool}_w(\chi'); \text{AvgPool}_w(\chi')]\) where \(\chi' \in \mathbb{R}^{W \times H \times C}\) is a 90 degree anti-clockwise rotation along the \(H\) axis from the output of the previous convolutional layer \(\chi \in \mathbb{R}^{C \times H \times W}\) and \(\chi_1 \in \mathbb{R}^{2 \times H \times C}\) is the output of a Z-Pool operation. \(\chi_1\) then passed through a standard 2D convolutional layer followed by sigmoid activation \(\sigma\) to get attention weights for \(\chi_1\). It would finally rotate back to match the original shape of \(\chi\) after applying the attention weights. These steps can be represented by the following: \(y_1 = r(\sigma(\text{CNN}_1(\chi_1)))\) where \(r\) is the rotation operation to retain the original shape of input. Similarly, \(y_2, y_3\) are obtained from the last two branches by measuring the interactions between dimensions \(W\) and \(C\) and between dimensions \(W\) and \(H\), respectively. Note that the last branch is similar to the spatial attention in CBAM\(^ {23}\), and it requires no rotation. The refined input \(y\) is represented by averaging outputs from three branches: \(y = \frac{1}{3}(y_1 + y_2 + y_3)\).

2.2 Training Strategy of ChexRadiNet

ChexRadiNet adopts an end-to-end multi-task training scheme. Each epoch consists of two tasks. In the first task (Branch I), we use the whole image to fine-tune the ResNet + Triplet Attention network pre-trained on ImageNet. During this process, we feed the generated masks into the radiomics extraction block to get radiomic features. In the second task (Branch II), we use radiomic features as regularization to further fine-tune the whole model. In each epoch, we use the model with the highest AUC on the validation set for testing.

3 Experiments

3.1 Datasets

For the abnormality classification task, we evaluated the ChexRadiNet framework using the NIH Chest X-ray\(^ {4}\), CheXpert\(^ {6}\), and MIMIC-CXR\(^ {7}\) datasets (Table 1). The Chest X-ray dataset contains 112,120 X-ray images collected from 30,805 patients. The disease labels were extracted from radiological reports with Natural Language Processing tools\(^ {24}\). There are 15 classes, one for “No findings” and 14 diseases: Atelectasis, Cardiomegaly, Consolidation, Edema, Effusion, Emphysema, Fibrosis, Hernia, Infiltration, Mass, Nodule, Pleural thickening, Pneumonia, and Pneumothorax. The disease labels are expected to have above 90% accuracy. In addition, the Chest X-ray dataset includes 984 bounding boxes for 8 types of chest diseases annotated for 880 images by radiologists.

CheXpert dataset is another large-scale public chest X-ray dataset currently available, which contains 224,316 X-ray scans of 65,240 patients. This dataset was labeled for the presence of 14 observations, including 12 common thoracic pathologies. Each observation can be assigned to either positive (1), negative (0), or uncertain (-1). To simplify the task, we choose to ignore all the uncertain samples. In addition, to compare with previous literature, we follow the same evaluation protocol over 5 observations: Atelectasis, Cardiomegaly, Consolidation, Edema, and Pleural Effusion.

MIMIC-CXR is also a large-scale CXR dataset, which contains 377,110 chest X-rays associated with 227,827 imaging studies. Images are provided with 13 labels. Similar to CheXpert, each label can be assigned to either positive (1), negative (0), or uncertain (-1).

| Table 1: Descriptions of the datasets. |
|-------------------------------|-----|-------|
| Datasets          | Patients | Chest X-rays |
| NIH Chest X-ray   | 30,805   | 112,120 |
| CheXpert          | 65,240   | 224,316 |
| MIMIC-CXR         | 227,827  | 377,110 |
3.2 Evaluation metrics and experimental settings

For the abnormality detection task, we randomly split each dataset into training (70%), validation (10%), and test (20%) sets. Note that there is no patient overlap between the sets. We use AUC scores, the area under the ROC curve, to measure the disease identification accuracy. A higher AUC score indicates better performance.

For the abnormality localization task, following the work of Li et al\(^1\), we only consider 8 diseases for the evaluation of mask generation because only eight types of diseases are provided with bounding boxes in the NIH Chest X-ray dataset. We use intersection over union (IoU) to evaluate the predicted disease regions against the ground truth bounding boxes.

We use ResNet-50 as the backbone model. We set the batch size as 256 and train the model for 20 epochs. The model is optimized using the stochastic gradient descent (SGD) optimizer with a learning rate of 0.1 and decay the learning rate by 0.1 every 5 epochs of training. We trained our model on AWS with 16 Nvidia K80 GPUs. The model is implemented in PyTorch.

3.3 Results

3.3.1 Disease classification

Table 2 shows the AUC of each class and a mean AUC across the 14 chest diseases. We used ResNet-50 pre-trained on ImageNet as the backbone. Our ChexRadiNet outperforms other models in terms of mean AUC. For every single class, our proposed framework is better than all other models except with DensNet-121 for Fibrosis, Hernia, Mass, Nodule, Pneumonia, and Pneumothorax. Possible reasons can be that Rajpurkar et al’s backbone is much deeper than our ResNet-50\(^3\), which enables it to capture more discriminative features than our ResNet-50. In addition, “Mass” and “Nodule” parts are small and hard to detect. For “Fibrosis” and “Hernia,” they are not annotated with bounding boxes and diffuse, and thus we cannot apply the weakly-supervised learning with radiomic features.

<table>
<thead>
<tr>
<th>Method</th>
<th>Atelectasis</th>
<th>Cardiomegaly</th>
<th>Consolidation</th>
<th>Edema</th>
<th>Effusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wang et al., 2017(^4)</td>
<td>0.716</td>
<td>0.807</td>
<td>0.708</td>
<td>0.835</td>
<td>0.784</td>
</tr>
<tr>
<td>Wang et al., 2018(^5)</td>
<td>0.732</td>
<td>0.844</td>
<td>0.701</td>
<td>0.829</td>
<td>0.793</td>
</tr>
<tr>
<td>Yao et al., 2018(^9)</td>
<td>0.772</td>
<td>0.904</td>
<td>0.788</td>
<td>0.882</td>
<td>0.859</td>
</tr>
<tr>
<td>Rajpurkar et al., 2017(^3)</td>
<td>0.821</td>
<td>0.905</td>
<td>0.794</td>
<td>0.893</td>
<td>0.883</td>
</tr>
<tr>
<td>Kumar et al., 2017(^25)</td>
<td>0.762</td>
<td>0.913</td>
<td>0.784</td>
<td>0.888</td>
<td>0.864</td>
</tr>
<tr>
<td>ChexRadiNet</td>
<td><strong>0.831</strong></td>
<td><strong>0.934</strong></td>
<td><strong>0.817</strong></td>
<td><strong>0.906</strong></td>
<td><strong>0.892</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Emphysema</th>
<th>Fibrosis</th>
<th>Hernia</th>
<th>Infiltration</th>
<th>Mass</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wang et al., 2017(^4)</td>
<td>0.815</td>
<td>0.769</td>
<td>0.767</td>
<td>0.609</td>
<td>0.706</td>
</tr>
<tr>
<td>Wang et al., 2018(^5)</td>
<td>0.865</td>
<td>0.796</td>
<td>0.876</td>
<td>0.666</td>
<td>0.725</td>
</tr>
<tr>
<td>Yao et al., 2018(^9)</td>
<td>0.829</td>
<td>0.767</td>
<td>0.914</td>
<td>0.695</td>
<td>0.792</td>
</tr>
<tr>
<td>Rajpurkar et al., 2017(^3)</td>
<td><strong>0.926</strong></td>
<td><strong>0.804</strong></td>
<td><strong>0.939</strong></td>
<td>0.720</td>
<td><strong>0.862</strong></td>
</tr>
<tr>
<td>Kumar et al., 2017(^25)</td>
<td>0.898</td>
<td>0.756</td>
<td>0.802</td>
<td>0.692</td>
<td>0.750</td>
</tr>
<tr>
<td>ChexRadiNet</td>
<td>0.925</td>
<td>0.798</td>
<td>0.882</td>
<td><strong>0.734</strong></td>
<td>0.846</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Nodule</th>
<th>Pleural Thickening</th>
<th>Pneumonia</th>
<th>Pneumothorax</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wang et al., 2017(^4)</td>
<td>0.671</td>
<td>0.708</td>
<td>0.633</td>
<td>0.806</td>
<td>0.738</td>
</tr>
<tr>
<td>Wang et al., 2018(^5)</td>
<td>0.685</td>
<td>0.735</td>
<td>0.720</td>
<td>0.847</td>
<td>0.772</td>
</tr>
<tr>
<td>Yao et al., 2018(^9)</td>
<td>0.717</td>
<td>0.765</td>
<td>0.713</td>
<td>0.841</td>
<td>0.803</td>
</tr>
<tr>
<td>Rajpurkar et al., 2017(^3)</td>
<td><strong>0.777</strong></td>
<td><strong>0.814</strong></td>
<td><strong>0.763</strong></td>
<td><strong>0.893</strong></td>
<td><strong>0.842</strong></td>
</tr>
<tr>
<td>Kumar et al., 2017(^25)</td>
<td>0.666</td>
<td>0.774</td>
<td>0.715</td>
<td>0.859</td>
<td>0.795</td>
</tr>
<tr>
<td>ChexRadiNet</td>
<td>0.748</td>
<td><strong>0.867</strong></td>
<td>0.737</td>
<td>0.889</td>
<td><strong>0.843</strong></td>
</tr>
</tbody>
</table>
Table 3: Disease localization under varying IoU on the NIH Chest X-ray dataset. Please note that since our model doesn’t use any ground truth bounding box information, to fairly evaluate the performance of our model, we only consider the previous methods’ results under the same setting, therefore, for the case T(IoU)=0.1, we have two baselines, but for the rest cases, we only have one baseline.

<table>
<thead>
<tr>
<th>T(IoU)</th>
<th>Model</th>
<th>Atelectasis</th>
<th>Cardiomegaly</th>
<th>Effusion</th>
<th>Infiltration</th>
<th>Mass</th>
<th>Nodule</th>
<th>Pneumonia</th>
<th>Pneumothorax</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>Wang et al., 2017⁴</td>
<td>0.69</td>
<td>0.94</td>
<td>0.66</td>
<td>0.71</td>
<td>0.40</td>
<td>0.14</td>
<td>0.63</td>
<td>0.38</td>
<td>0.569</td>
</tr>
<tr>
<td></td>
<td>Li et al., 2018¹</td>
<td>0.63</td>
<td>0.89</td>
<td>0.78</td>
<td>0.91</td>
<td>0.70</td>
<td>0.29</td>
<td>0.31</td>
<td>0.44</td>
<td>0.619</td>
</tr>
<tr>
<td></td>
<td>ChexRadiNet</td>
<td>0.72</td>
<td>0.96</td>
<td>0.81</td>
<td>0.88</td>
<td>0.67</td>
<td>0.33</td>
<td>0.59</td>
<td>0.47</td>
<td>0.679</td>
</tr>
<tr>
<td>0.2</td>
<td>Wang et al., 2017⁴</td>
<td>0.47</td>
<td>0.68</td>
<td>0.45</td>
<td>0.48</td>
<td>0.26</td>
<td>0.05</td>
<td>0.35</td>
<td>0.23</td>
<td>0.371</td>
</tr>
<tr>
<td></td>
<td>ChexRadiNet</td>
<td>0.49</td>
<td>0.84</td>
<td>0.62</td>
<td>0.54</td>
<td>0.46</td>
<td>0.21</td>
<td>0.43</td>
<td>0.39</td>
<td>0.498</td>
</tr>
<tr>
<td>0.3</td>
<td>Wang et al., 2017⁴</td>
<td>0.24</td>
<td>0.46</td>
<td>0.30</td>
<td>0.28</td>
<td>0.15</td>
<td>0.04</td>
<td>0.17</td>
<td>0.13</td>
<td>0.221</td>
</tr>
<tr>
<td></td>
<td>ChexRadiNet</td>
<td>0.28</td>
<td>0.73</td>
<td>0.54</td>
<td>0.43</td>
<td>0.38</td>
<td>0.15</td>
<td>0.35</td>
<td>0.32</td>
<td>0.398</td>
</tr>
<tr>
<td>0.4</td>
<td>Wang et al., 2017⁴</td>
<td>0.09</td>
<td>0.28</td>
<td>0.20</td>
<td>0.12</td>
<td>0.07</td>
<td>0.01</td>
<td>0.08</td>
<td>0.07</td>
<td>0.115</td>
</tr>
<tr>
<td></td>
<td>ChexRadiNet</td>
<td>0.17</td>
<td>0.65</td>
<td>0.42</td>
<td>0.32</td>
<td>0.29</td>
<td>0.09</td>
<td>0.21</td>
<td>0.19</td>
<td>0.293</td>
</tr>
<tr>
<td>0.5</td>
<td>Wang et al., 2017⁴</td>
<td>0.05</td>
<td>0.18</td>
<td>0.11</td>
<td>0.07</td>
<td>0.01</td>
<td>0.01</td>
<td>0.03</td>
<td>0.03</td>
<td>0.061</td>
</tr>
<tr>
<td></td>
<td>ChexRadiNet</td>
<td>0.11</td>
<td>0.59</td>
<td>0.29</td>
<td>0.15</td>
<td>0.12</td>
<td>0.07</td>
<td>0.14</td>
<td>0.08</td>
<td>0.194</td>
</tr>
<tr>
<td>0.6</td>
<td>Wang et al., 2017⁴</td>
<td>0.02</td>
<td>0.08</td>
<td>0.05</td>
<td>0.02</td>
<td>0.00</td>
<td>0.01</td>
<td>0.02</td>
<td>0.03</td>
<td>0.029</td>
</tr>
<tr>
<td></td>
<td>ChexRadiNet</td>
<td>0.06</td>
<td>0.37</td>
<td>0.09</td>
<td>0.06</td>
<td>0.08</td>
<td>0.04</td>
<td>0.05</td>
<td>0.05</td>
<td>0.100</td>
</tr>
<tr>
<td>0.7</td>
<td>Wang et al., 2017⁴</td>
<td>0.01</td>
<td>0.03</td>
<td>0.02</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.01</td>
<td>0.02</td>
<td>0.011</td>
</tr>
<tr>
<td></td>
<td>ChexRadiNet</td>
<td>0.02</td>
<td>0.21</td>
<td>0.04</td>
<td>0.02</td>
<td>0.07</td>
<td>0.01</td>
<td>0.03</td>
<td>0.04</td>
<td>0.055</td>
</tr>
</tbody>
</table>

3.3.2 Disease localization

We compare our disease localization accuracy under varying IoU to other state-of-the-art models, shown in Table 3. Our model predicts well not only for easy tasks but also for hard tasks like localizing “Mass” and “Nodule”, where the disease localization is within a small area. When the IoU is set to 0.1, our model outperforms other models in terms of Atelectasis, Cardiomegaly, Effusion, and Pneumothorax. As the IoU threshold increases, our framework is superior to other models in terms of better accuracy and maintains great performance. For instance, when IoU is set to 0.3, our result for “Cardiomegaly” is 0.73 while the reference model is only 0.46. We get more than 0.15 accuracy improvement for Effusion, Infiltration, Mass, Pneumonia, and Pneumothorax. When IoU is set to 0.5, our result for “Cardiomegaly” is still as high as 0.59 while the reference model drops to barely 0.18.

Following Li et al.¹¹, we prefer a higher IoU threshold, i.e., IoU = 0.7, for disease localization because we expect high-accuracy disease localization application in clinical use. To this end, the method we proposed is superior to the baseline by a large margin.

Please note that for some diseases, e.g., Pneumonia and Infiltration, the localization of disease can appear in multiple places while only one bounding box is provided for each image. Thus, it is reasonable that our model doesn’t align well with the ground truth when the threshold is as small as 0.1, especially for Pneumonia and infiltration. Overall, our model outperforms the reference models for all IoU thresholds except for T(IoU)=0.1 (probably because ground truth has missing annotation while ours does not).

4 Discussion

4.1 Ablation study

We conducted an ablation study to demonstrate the performance of radiomics on NIH Chest X-ray (Table 4), CheXpert (Table 5), and MIMIC-CXR (Table 6). We tried ResNet50+Triplet Attention without radiomic features. Table 4 shows that AUC will drop significantly when not using radiomic features. We observe the same trend in the other two datasets. This demonstrates that it is beneficial to include radiomic features.

We also report results of ChexxRadiNet using ResNet-18, a relevant small network, as a backbone. Table 7 shows the results with and without using the radiomic features in three datasets. We observe the AUCs drop significantly when
not using radiomic features in all cases. This suggests that the generalizability of our proposed method in smaller networks. In addition, the ResNet-18 version still performs better than other models in Table 2 except Rajpurkar et al. It indicates the superior of our proposed method for using radiomic features.

<table>
<thead>
<tr>
<th>Method</th>
<th>Atelectasis</th>
<th>Cardiomegaly</th>
<th>Consolidation</th>
<th>Edema</th>
<th>Effusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>w/o radiomics</td>
<td>0.751</td>
<td>0.850</td>
<td>0.777</td>
<td>0.867</td>
<td>0.833</td>
</tr>
<tr>
<td>ChexRadiNet</td>
<td>0.831</td>
<td>0.934</td>
<td>0.817</td>
<td>0.906</td>
<td>0.892</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Emphysema</th>
<th>Fibrosis</th>
<th>Hernia</th>
<th>Infiltration</th>
<th>Mass</th>
</tr>
</thead>
<tbody>
<tr>
<td>w/o radiomics</td>
<td>0.783</td>
<td>0.733</td>
<td>0.804</td>
<td>0.670</td>
<td>0.694</td>
</tr>
<tr>
<td>ChexRadiNet</td>
<td>0.925</td>
<td>0.798</td>
<td>0.882</td>
<td>0.734</td>
<td>0.846</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Nodule</th>
<th>Pleural Thickening</th>
<th>Pneumonia</th>
<th>Pneumothorax</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>w/o radiomics</td>
<td>0.643</td>
<td>0.699</td>
<td>0.700</td>
<td>0.792</td>
<td>0.757</td>
</tr>
<tr>
<td>ChexRadiNet</td>
<td>0.748</td>
<td>0.867</td>
<td>0.737</td>
<td>0.889</td>
<td>0.842</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Atelectasis</th>
<th>Cardiomegaly</th>
<th>Consolidation</th>
<th>Edema</th>
<th>Pleural Effusion</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>w/o radiomics</td>
<td>0.781</td>
<td>0.813</td>
<td>0.893</td>
<td>0.918</td>
<td>0.921</td>
<td>0.865</td>
</tr>
<tr>
<td>ChexRadiNet</td>
<td>0.831</td>
<td>0.848</td>
<td>0.920</td>
<td>0.930</td>
<td>0.921</td>
<td>0.890</td>
</tr>
</tbody>
</table>

| Method                  | Atelectasis | Cardiomegaly | Consolidation | Edema   | Enlarged Card.  |
|-------------------------|-------------|--------------|---------------|---------|----------------|-------|
| w/o radiomics           | 0.841       | 0.824        | 0.859         | 0.906   | 0.748           | 0.767|
| ChexRadiNet             | 0.851       | 0.831        | 0.866         | 0.900   |                  |       |

<table>
<thead>
<tr>
<th>Method</th>
<th>Pneumothorax</th>
<th>Pleural Other</th>
<th>Support Devices</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>w/o radiomics</td>
<td>0.909</td>
<td>0.850</td>
<td>0.931</td>
<td>0.832</td>
</tr>
<tr>
<td>ChexRadiNet</td>
<td>0.919</td>
<td>0.909</td>
<td>0.937</td>
<td>0.854</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>NIH Chest X-ray</th>
<th>CheXpert</th>
<th>MIMIC-CXR</th>
</tr>
</thead>
<tbody>
<tr>
<td>w/o radiomics</td>
<td>0.749</td>
<td>0.854</td>
<td>0.822</td>
</tr>
<tr>
<td>ChexRadiNet (ResNet-18)</td>
<td>0.810</td>
<td>0.883</td>
<td>0.837</td>
</tr>
</tbody>
</table>

4.2 Qualitative analysis

Figure 2 shows the attention map of our model against the ground truth bounding boxes. The visualization provides better explainability of our model. In Figure 2 we visualized our results for Cardiomegaly, Mass, and Pneumonia. Cardiomegaly is considered to be present if the cardiothoracic rate is larger than 50% (cardiothoracic Ratio equals “Maximum horizontal cardiac width” over “Maximum horizontal thoracic width”), which means an enlarged heart. The 2nd image in the 1st row as well as the 2nd image in the 2nd row in Figure 2 shows that our model successfully detects cardiomegaly, an enlarged heart, perfectly, and aligned with the yellow bounding box well.

A lung mass is an abnormal spot in the lungs that is more than 3 centimeters. Our results (4th images in the 1st and 2nd rows), although focusing on larger areas, can capture some clues of lung mass.
Figure 2: Visualization of the disease localization on the test images with ChexRadiNet and ground truth bounding boxes. The attention maps are generated from the final output tensor and overlapped on the original radiology images. The left image in each pair is the chest X-ray image and the right one is the generated attention map and the ground truth (in the yellow box).

Note that in the chest X-ray 14 dataset, only one bounding box is annotated for one disease image. Though some patients are diagnosed with several diseases, only the most important disease is annotated on the radiology image. This means that ground truth has missing annotations (shown by Pneumonia). Pneumonia inflames the air sacs in one or both lungs. For Pneumonia detection, radiologists will look for white spots in the lungs. For the 6th image in the 2nd row, both lungs are infected and white spots are shown in both lungs. However, the bounding box of the 6th image only annotates the right lung while our model successfully localizes Pneumonia for both lungs.

Overall, our results show that the predicted disease localizations have a great alignment with the ground truth and can even serve as a supplement to the ground truth.

5 Conclusion

We propose a framework that jointly learns radiomic features and predicts 14 thoracic diseases. We evaluated our model on three publicly available corpora. We showed that both our disease identification and localization outperform state-of-the-art models in the quantitative and qualitative analysis.

Our proposed framework has two main limitations. First, chest X-rays are very different from natural images, but we rely on deep learning models (ResNet) that work better on natural images. Second, the robustness of radiomic features relies on the accuracy of bounding boxes, in our work, the bounding boxes are generated by heatmaps. It is not guaranteed that the generated heatmaps are always good and accurate. Our future work will continue to solve these two limitations.

Automatically generating correct bounding boxes can be a milestone to push the agenda for AI-driven medical imaging diagnosis. It can abruptly increase the annotated medical images at a much lower cost so that better CNN models can be trained, therefore better diagnosis models can be obtained. Bounding boxes can increase the interpretability of AI solutions by locating the abnormalities as the visual evidence in medical images, which can build trust between doctors and patients.

Acknowledgment

This work is supported by Amazon Machine Learning Research Award 2020. It also was supported by the National Library of Medicine under Award No. 4R00LM013001.
References


11. Zhe Li, Chong Wang, Mei Han, Yuan Xue, Wei Wei, Li-Jia Li, and Li Fei-Fei. Thoracic disease identification and localization with limited supervision. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 8290–8299, 2018.


A Federated Mining Approach on Predicting Diabetes-Related Complications: Demonstration Using Real-World Clinical Data

Humayera Islam, MS1,4, Abu Mosa, PhD, FAMIA*1,2,3,4

1Institute for Data Science and Informatics; 2Department of Health Management and Informatics; 3Department of Electrical Engineering and Computer Science; 4Center for Biomedical Informatics; University of Missouri School of Medicine, Columbia, Missouri

*Corresponding Author. Email: mosaa@health.missouri.edu

Abstract

Chronic diabetes can lead to microvascular complications, including diabetic eye disease, diabetic kidney disease, and diabetic neuropathy. However, the long-term complications often remain undetected at the early stages of diagnosis. Developing a machine learning model to identify the patients at high risk of developing diabetes-related complications can help design better treatment interventions. Building robust machine learning models require large datasets which further requires sharing data among different healthcare systems, hence, involving privacy and confidentiality concerns. The main objective of this study is to design a decentralized privacy-protected federated learning architecture that can deliver comparable performance to centralized learning. We demonstrate the potential of adopting federated learning to address the challenges such as class-imbalance in using real-world clinical data. In all our experiments, federated learning showed comparable performance to the gold-standard of centralized learning, and applying class balancing techniques improved performance across all cohorts.

Introduction

Diseases such as chronic diabetes have evidence of engendering other fatal long-term comorbid complications. For instance, prolonged uncontrolled diabetes can lead to microvascular complications, including diabetic eye disease, diabetic kidney disease, and diabetic neuropathy1-3. Diabetes-related complications resulted in 16 million emergency department visits and 7.8 million hospitalizations, as estimated in 2016, contributing to an increasing burden on the US healthcare system. Diabetes is also the leading cause of end-stage kidney disease, with a crude prevalence of 38% and new cases of vision disability with a crude rate of 11.7%3,4.

However, the long-term complications often remain undetected at the early stages of diagnosis, and most of the treatment plans for addressing the complications are reactive rather than proactive. Developing a state-of-the-art prediction model that can learn from the patient-related factors to identify the patients at high risk of developing diabetes-related complications can help design better treatment interventions. Hence, it can play an essential role in minimizing costs related to hospitalizations, medications, and treatment procedures ensuing from those complications.

Digitizing healthcare data in electronic health records (EHR) evolved into a rich source of patient health history, thus generating more opportunities to innovate data-driven tools and techniques to improve the availability and accuracy of medical services5. Potential utilization of machine learning in predicting health outcomes can involve incorporating large amounts of data generated by independent health systems6,7. However, developing centralized algorithms for centralized data repositories raises privacy, confidentiality, and regulatory concerns such as the HIPAA and HITECH Acts8. For instance, incidences of diabetes-related complications are rare events among the diabetes population. Thus, it is imperative to accumulate data from different healthcare systems to build reliable and robust predictive models using a large representation of the population for such rare complications. To address the persistent concerns related to data-sharing among the healthcare systems, we present a federated learning-based framework that can consolidate predictive models without using central repositories of the actual data itself.

In recent times, the concepts of federated learning have been seamlessly adapted in the healthcare domain to address the privacy concerns on sharing potential patient information among the healthcare systems. A decentralized prediction engine can potentially use data stored from independent health systems without ethical or legal circumstances. This study is the first implementation of federated learning in predicting diabetes-related complications using real-world clinical data to the best of our knowledge. Moreover, predictions of diabetes-related complications using federated learning encounter significant challenges from the inherent class-imbalance characteristics of real-world clinical data. The main objective of this study is to design a decentralized privacy-protected federated learning
architecture that can deliver comparable performance to centralized learning. We demonstrate the potential of adopting federated learning to address the challenges in using real-world clinical data.

The key contributions of our study include (i) identifying a diabetes population and three cohorts related to diabetic eye disease, diabetic kidney disease, and diabetic retinopathy from Health Facts EMR data using a structured framework, (ii) implementing a decentralized privacy-controlled federated learning architecture that can utilize the federated datasets from different healthcare systems to predict the complication in each cohort without sharing data among themselves, (iii) performing sampling techniques to address the class-imbalance characteristic of the datasets in the federated learning architecture, and (iv) comparing the performance of federated learning to centralized learning in different sampling conditions.

Background
Federated learning involves iteratively analyzing separate databases and sharing only mathematical parameters (metadata), but not the actual data itself that might reveal potential patient identifiers. Federated learning mechanisms were initially more popular in image classification and enhancing wireless communication systems. The recent adaptation of federated learning models in the healthcare domain includes predictions on healthcare outcomes such as mortality, hospital stay-time for ICU patients, hospitalization for cardiac events, dyspnea, adverse drug reactions. However, most of the implementation of federated learning in predicting healthcare outcomes utilized small datasets and partitioned the data hypothetically (randomly) to mimic the inherent characteristics of real-world data. In this study, we utilize the natural partitions of the Health Facts data using the information of the healthcare systems for each patient data to demonstrate our framework.

In most federated learning applications in the healthcare domain, classification algorithms such as logistic regression, artificial neural network, multi-layer perceptron, support vector machines, random forest were used to build federated predictive models. Existing literature in predicting diabetic retinopathy (eye disease), neuropathy (peripheral nerve disorder), and nephropathy (kidney disease) involve centralized machine learning algorithms using small-size datasets from the US population with limited cases of complications and limited patient information. For our study, we implemented three machine learning models, including logistic regression, 2-layer multiple perceptrons, and 3-layer multi-perceptron in federated learning architecture for binary classification of the incidence of three diabetes-related complications affecting eyes, kidneys, and peripheral nerves, respectively.

Methods
Data Source
We have used Cerner’s “Health Facts EMR Data” as our data source in this study. Health Facts is a de-identified electronic health records relational database consolidated from over 90 healthcare systems in the US between 2000 and 2016 in which Cerner has a data use agreement. Health Facts database contains patient-level attributes, such as demographics, encounters, diagnoses, lab results, procedures, prescription orders, and other clinical observations on 69 million unique patients. We identified the diabetes population, defined three cohorts for diabetes-related complications, and extracted patient-level comorbid features for our cohorts using a detailed pipeline on this dataset.

Diabetes Population Selection
We identified the diabetes population in our study using the SUPREME-DM (Surveillance, PREvention, and ManagEment of Diabetes Mellitus algorithm) based on eight criteria (Table 1). Six criteria were based on lab results, while two criteria were based on International Classification of Disease (ICD-9 and ICD-10) diagnosis codes related to inpatient and outpatient encounters. Patients satisfying at least one criterion or more were selected as the diabetes population. Only patients aged 18 or above at the first encounter were included.

Selection of Cohorts with Diabetes-Related Complications
The diabetes-related complications among the selected diabetes population were identified using International Classification of Disease (ICD-9 and ICD-10) diagnosis codes. We selected three cohorts, including diabetic eye disease (EDD), diabetic kidney disease (KDD), and diabetic neuropathy (ND), from the selected diabetes population. Figure 1 shows the flow diagram of the selection of cohorts with ICD codes for each of the diabetes-related complications.
Table 1: We used the following eight criteria to select the diabetes population from Health Facts. The first six criteria were based on lab results, while the last two were based on ICD-9 and ICD-10 diagnosis codes. The thresholds for the lab tests are chosen based on the SUPREME-DM. Patients satisfying at least one criterion were selected as the diabetes population.

<table>
<thead>
<tr>
<th>Criterion (at least one)</th>
<th>Value</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>HbA1c</td>
<td>≥ 2 and ≥ 6.5%</td>
<td>Tests must be within two years apart</td>
</tr>
<tr>
<td>Fasting plasma glucose</td>
<td>≥ 2 and ≥ 126 mg/dL</td>
<td></td>
</tr>
<tr>
<td>Random plasma glucose</td>
<td>≥ 2 and ≥ 200 mg/dL</td>
<td></td>
</tr>
<tr>
<td>Random plasma plus fasting glucose</td>
<td>1 at ≥200 mg/dL and 1 at ≥126 mg/dL</td>
<td></td>
</tr>
<tr>
<td>HbA1c plus fasting glucose</td>
<td>1 at ≥6.5% and 1 at ≥126 mg/dL</td>
<td></td>
</tr>
<tr>
<td>HbA1c plus random plasma glucose</td>
<td>1 at ≥6.5% and 1 at ≥200 mg/dL</td>
<td></td>
</tr>
<tr>
<td>Inpatient discharge diagnosis</td>
<td>≥1 inpatient visit with one of the following diagnosis codes primary: 250.x, 357.2, 366.41, 362.01–362.07, and E08.x-E13.x</td>
<td>Primary diagnosis code only</td>
</tr>
<tr>
<td>Outpatient visit diagnosis</td>
<td>≥2 outpatient visits with one of the following diagnosis codes 250.x, 357.2, 366.41, 362.01–362.07, and E08.x-E13.x</td>
<td>Visits must occur on separate days (ambulatory visits only)</td>
</tr>
</tbody>
</table>

Many studies have shown that patients diagnosed with diabetes show higher chances of developing these complications after five years of chronic exposure\textsuperscript{20–22}. Patients diagnosed with any of the above complications who had at least five years of chronic diabetes with total encounters between 25 and 500 were selected as “cases”. For the “non-cases”, patients with at least seven years of chronic diabetes with no diagnosis of any of the complications and total encounters between 25-500 were selected. The patient-related comorbid features were extracted using the diagnosis table.

Table 2: The summary statistics for each of the three cohorts among the diabetes population are shown below. We excluded the healthcare systems with less than 100 cases from the dataset, which reduced the number of cases in each cohort, as shown in the following table.

<table>
<thead>
<tr>
<th>Diabetic Population</th>
<th>Diabetic Eye Disease Cohort</th>
<th>Diabetic Kidney Disease Cohort</th>
<th>Diabetic Neuropathy Cohort</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Patients</td>
<td>102,876 (100)</td>
<td>10,599 (10.3)</td>
<td>17,455 (17.1)</td>
</tr>
<tr>
<td>Number of Cases (After Exclusion)</td>
<td>-</td>
<td>9,686 (9.4)</td>
<td>16,727 (16.3)</td>
</tr>
<tr>
<td>Class-Ratio (Cases: No Cases)</td>
<td>13:100</td>
<td>22:100</td>
<td>32:100</td>
</tr>
<tr>
<td>Number of Independent Healthcare Systems</td>
<td>70</td>
<td>22</td>
<td>31</td>
</tr>
</tbody>
</table>

The diagnosis ICD codes were mapped into the Clinical Classification Software (CCS) tool developed by Healthcare Cost and Utilization Project (HCUP). We used 283 CCS coded features consolidated from the individual ICD-9 and ICD-10 codes by grouping the ICD codes into clinically similar entities, which are used as predictors for the experiments. Table 2 shows the summary statistics for the selected three cohorts of diabetes-related complications.

Decentralized Data Architecture
To demonstrate the method of privacy-preserved federated learning using the three selected cohorts, we utilized the hospital system identifiers for each patient in each cohort. The variable “ALT_HEALTH_SYSTEM_ID” was used to extract the hospital identifiers from the “hs_d_hospital” data table for each “PATIENT_ID”. The patient IDs were mapped to “PATIENT_SK”, which were linked to the CCS table of patient features for each cohort. Table 2 shows the number of hospital systems in each of the cohorts. We distributed the data for each cohort into small datasets for each hospital system to facilitate our federated learning architecture, as discussed below. The healthcare systems with less than 100 cases were discarded from the analysis.

Figure 1: Flow diagram showing the cohort selection for diabetic eye disease, diabetic kidney disease, and diabetic neuropathy from the diabetes population using the shown diagnosis codes.
Centralized Predictive Model

A centralized predictive model approach was implemented to represent the characteristics of a centralized data warehouse gathering data from multiple sites. We considered a general binary classification of the diabetes-related complications in this study. Each primary response variable, including diabetic eye disease, diabetic kidney disease, and diabetic neuropathy, is categorized into binary responses, such as cases vs. non-cases. The CCS comorbid features of each patient in each cohort are used to predict the complication as binary responses. For this purpose, we considered logistic regression, 2-layer multiple layer perceptron, and 3-layer multiple layer perceptron models. The machine learning models were trained using 70% of the complete dataset, while 30% were used for testing and validation. The algorithms were implemented using the sci-kit learn module from Python 3.6. The centralized approach to predict each complication from each cohort will act as a benchmark to compare with the federated learning approach.

Federated Predictive Model

The federated learning approach utilized the partitioned data from each cohort to predict the binary cases and non-cases for each complication. We used sci-kit learn, NumPy, pandas for the machine learning tasks, and TensorFlow to create the federated mining pipeline. To demonstrate this approach, separate training and testing datasets were created for each disjointed partitioned dataset, i.e., each healthcare system in each cohort. The labels from training data for each cohort were 1-hot-encoded using the LabelBinarizer() object from sk-learn. The features were then transformed into TensorFlow data objects. We used binary cross-entropy as loss function and stochastic gradient descent as the optimizer to compile three models: logistic regression, a 2-layer multi-perceptron, and a 3-layer multi-perceptron. Since the data in our study are horizontally partitioned, local model parameters will contribute to the global model weighted by the proportion of data points from each participating healthcare system.

We developed the training module for federated learning using the federated averaging algorithm from McMahan (2017)\(^8\). Algorithm 1 shows the two main functions used to train the federated models. At first, the global model weights are initialized, which serve as the initial weights for all local models. Using these initial weights, the local models for each healthcare system \(k = 1, ..., K\) are trained to obtain the updated weights. The updated weights are then scaled by a factor \(\frac{N_k}{N}\), which is the proportion of data for \(k^{th}\) Healthcare system and summed to obtain a new set of weights. These averaged weights (as shown in line 9 from Algorithm 1) are then used to update the global model. The process continues until \(T\) rounds of aggregating local model weights to update the global model. Thus, a federated learning mechanism only relies on sharing the weights of the local models without sharing any raw data from the individual healthcare systems. We repeat these steps to predict the three complications in the three cohorts separately. Figure 3 shows the system design of federated learning deployed in our study.

---

Algorithm 1: Federated Learning Model for Predicting Diabetes-Related Complications. The \(K\) healthcare systems are indexed by \(k\); \(E\) is the number of epochs, \(b\) is the mini-batch size, \(\eta\) is the learning rate, and \(l(\omega; b)\) is the loss function.

1: function GLOBALMODELUPDATE  
2:     initialize \(\omega_0\)  
3:     for \(t = 1\) to \(T\);  
4:         for \(k = 1\) to \(K\);  
5:             \(\omega_{t+1}^k = \text{LOCALMODELUPDATE}(k, \omega_t)\)
6: \[ n_k = \sum n(k) \]
7: \[ N = \sum_{k=1}^{K} n_k \]
8: \[ \omega_{t+1} = \frac{n_k}{N} \cdot \omega_{t+1} \]

9: return \((\omega_{t+1})\) to update global model

10: function LOCALMODELUPDATE \((i, \omega)\) #Run on Hospital System \(k\)
11: for \(e = 1 \text{ to } E\):
12: for \(b \in \text{Batch}\):
13: \[ \omega = \omega - \eta \nabla l(\omega; b) \]
14: return \((\omega)\)

---

**Class-Imbalance Learning**

As shown in Figure 2, an imbalance between the cases and non-cases of the three diabetes-related complications among the diabetes population exists. Moreover, the federated datasets for the three cohorts are also subjected to unequal distribution of cases and non-cases. This is consistent with our expectations since the complications are rare for most patients diagnosed with diabetes. Classification algorithms applied in datasets with non-cases as the minority class is more likely to predict new observations in the majority class since they fail to characterize their imbalanced nature. Also, federated averaging would not account for the varying distribution of cases across the different healthcare systems. Thus, it is imperative to consider this inherent characteristic of class distribution in our model training. We applied sampling techniques, such as oversampling and undersampling, to address the class-imbalance attribute of the cohort datasets. Oversampling supplement the minority class, whereas undersampling randomly removes the majority class. For centralized learning, oversampling, undersampling, and no-sampling were performed on the training datasets for each cohort before model compilation. In the federated learning approach, similar sampling strategies were applied on the federated datasets for each cohort before the local model compilations. The module “resample” from sk-learn was used for executing the oversampling and undersampling techniques with “n_samples” parameter set to the class size for both over- and under-sampling.

**Experimentation and Evaluation Metrics**

The three machine learning models were run for each of the cohorts for the three sampling techniques including, no balancing, oversampling, and undersampling. In total, there were 27 experiments for the federated learning.
Architecture and 27 experiments for centralized learning. We combined the testing data from all sites within each cohort to evaluate the performance of the federated model. When the datasets are subjected to class-imbalance, F-1 scores are a more reliable measure compared to accuracy. Performance metrics, such as F-1 score, recall, and precision, were used to compare the performance of the federated learning to centralized learning.

R version 3.4.4 (R Foundation for Statistical Computing, Vienna, Austria) and Python 3.6 were used for data management and all computations were performed on a Mac Book Pro running macOS Catalina version 10.15.2 with 16GB of RAM.

Comparative Analysis
The performance measures across all the classifiers and datasets showed federated learning to exhibit comparable predictive performance with respect to centralized learning. As shown in Table 3, multilayer perceptron consistently performed well compared to logistic regression in predicting diabetic eye disease patients. For logistic regression, applying the undersampling improved the recall by 12% for both federated and centralized learning. The F-1 scores for both multi-layer perceptron models were improved by about 4% for federated learning and 13% for centralized learning. However, precision measures decreased by 20% for both oversampling and undersampling experiments for both the learning mechanisms. The difference between the precision scores between federated and centralized learning varied by 6-7%.

Table 3: Performance metrics (F-1 score, Precision, and Recall) for centralized learning and federated learning, obtained from three machine learning models: logistic regression, 2-layer multi-perceptron, 3-layer multi-perceptron for the diabetic eye disease cohort.

<table>
<thead>
<tr>
<th>Model</th>
<th>Method</th>
<th>Federated Learning</th>
<th>Centralized Learning</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR</td>
<td>Oversample</td>
<td>0.59</td>
<td>0.69</td>
</tr>
<tr>
<td>LR</td>
<td>Undersample</td>
<td>0.59</td>
<td>0.66</td>
</tr>
<tr>
<td>LR</td>
<td>No Balancing</td>
<td>0.62</td>
<td>0.75</td>
</tr>
<tr>
<td>MLP (10,1)</td>
<td>Oversample</td>
<td>0.63</td>
<td>0.71</td>
</tr>
<tr>
<td>MLP (10,1)</td>
<td>Undersample</td>
<td>0.59</td>
<td>0.61</td>
</tr>
<tr>
<td>MLP (10,1)</td>
<td>No Balancing</td>
<td>0.68</td>
<td>0.75</td>
</tr>
<tr>
<td>MLP (10,10,1)</td>
<td>Oversample</td>
<td>0.65</td>
<td>0.67</td>
</tr>
<tr>
<td>MLP (10,10,1)</td>
<td>Undersample</td>
<td>0.59</td>
<td>0.61</td>
</tr>
<tr>
<td>MLP (10,10,1)</td>
<td>No Balancing</td>
<td>0.69</td>
<td>0.66</td>
</tr>
</tbody>
</table>

All the classifiers from the two setups performed better in the diabetic kidney disease cohort, as shown in Table 4. Both oversampling and undersampling improved recall for logistic regression and multi-payer perceptron models. For both centralized and federated learning, recall improved by 20%. The maximum difference in recall between federated and centralized learning was within 3% for both the oversampling and undersampling cases. Maximum F-1 scores were obtained by the multi-layer perceptron models and were comparable for both federated and centralized learning. The maximum difference in F-1 scores between federated and centralized learning were about 4% when the balancing techniques were applied. When no balancing technique was implemented, the maximum difference in the performance metrics between the two learning setups is about 10%.

As shown in Table 5, for logistic regression applying sampling techniques improved the F-1 score by 6-7% for federated learning and about 1% for centralized learning. Recall improved by 34-35% in federated logistic regression, while a 26% increase is observed in centralized logistic regression. The maximum difference in F-1 score and recall between federated and centralized logistic regression is about 2%. In the 2-layer multi-perceptron model, sampling techniques improved F-1 score by 4-5% in federated learning, while an improvement of 6-7% was observed in centralized learning. Recall improved by 24% in federated and 31% in centralized learning setups upon implementing sampling techniques. The maximum difference between F-1 scores is about 3%, while only a 1% difference in the recall. The 3-layer multi-perceptron model is the best performing model in the diabetic neuropathy cohort. Oversampling improved recall by 10% for federated learning vs. centralized learning, while equivalent recall values were obtained for both. F-1 measure improved by 2% and equivalent precision measures were also obtained for
federated learning in this scenario. Among the three cohorts, federated learning with sampling performed consistently well in the diabetic neuropathy cohort.

**Table 4:** Performance metrics (F-1 score, Precision, and Recall) for centralized learning and federated learning, obtained from three machine learning models: logistic regression, 2-layer multi-perceptron, 3-layer multi-perceptron for the diabetic kidney disease cohort.

<table>
<thead>
<tr>
<th>Model</th>
<th>Method</th>
<th>Federated Learning</th>
<th>Centralized Learning</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR</td>
<td>Oversample</td>
<td>0.65 0.56 0.85</td>
<td>0.68 0.59 0.85</td>
</tr>
<tr>
<td>LR</td>
<td>Undersample</td>
<td>0.62 0.53 0.83</td>
<td>0.67 0.56 0.86</td>
</tr>
<tr>
<td>LR</td>
<td>No Balancing</td>
<td>0.56 0.72 0.5</td>
<td>0.64 0.76 0.6</td>
</tr>
<tr>
<td>MLP (10,1)</td>
<td>Oversample</td>
<td>0.67 0.58 0.85</td>
<td>0.66 0.56 0.84</td>
</tr>
<tr>
<td>MLP (10,1)</td>
<td>Undersample</td>
<td>0.65 0.56 0.84</td>
<td>0.65 0.54 0.87</td>
</tr>
<tr>
<td>MLP (10,1)</td>
<td>No Balancing</td>
<td>0.61 0.69 0.6</td>
<td>0.65 0.74 0.61</td>
</tr>
<tr>
<td>MLP(10,10,1)</td>
<td>Oversample</td>
<td>0.64 0.54 0.84</td>
<td>0.63 0.53 0.84</td>
</tr>
<tr>
<td>MLP(10,10,1)</td>
<td>Undersample</td>
<td>0.64 0.53 0.85</td>
<td>0.67 0.57 0.85</td>
</tr>
<tr>
<td>MLP(10,10,1)</td>
<td>No Balancing</td>
<td>0.63 0.64 0.66</td>
<td>0.65 0.74 0.61</td>
</tr>
</tbody>
</table>

**Discussion**

The use of federated learning can bring numerous opportunities to investigate rare clinical events by building decentralized models without exchanging direct raw data. In this paper, we developed a decentralized privacy-protected predictive classifier that can successfully predict diabetes-related complications such as retinopathy, neuropathy, and nephropathy. The federated architecture shares a common global model with the healthcare systems, which utilize their electronic health records to train locally. Updates from the local models are aggregated later to update the global model. In this process, no personal patient identifiers are shared. We used the Health Facts database to demonstrate the process, which is an information-enriched database for patient records. We used the unique identifiers for healthcare systems to partition the data into real federated data sets. We applied class balancing sampling techniques to address the challenge of low class-ratios of cases to non-cases. In all our experiments, federated learning showed comparable performance to the gold-standard of centralized learning.

Moreover, federated learning with 3-layer multi-perceptron model performed consistently better than its centralized counterparts in the diabetic neuropathy cohort. Interestingly, this cohort had fewer class-imbalance issues, hence greater number of cases of neuropathy in the dataset, which accounts for the consistent performance of federated learning in this case. Also, it is evident that applying the class balancing techniques reduced the gap between the measures of federated and centralized learning across all the cohorts. Thus, federated learning models are useful for the healthcare systems where data sharing is a major barrier for building machine-learning based clinical decision support systems.

**Table 5:** Performance metrics (F-1 score, Precision, and Recall) for centralized learning and federated learning, obtained from three machine learning models: logistic regression, 2-layer multi-perceptron, 3-layer multi-perceptron for the diabetic neuropathy cohort.

<table>
<thead>
<tr>
<th>Model</th>
<th>Method</th>
<th>Federated Learning</th>
<th>Centralized Learning</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR</td>
<td>Oversample</td>
<td>0.63 0.52 0.87</td>
<td>0.65 0.53 0.87</td>
</tr>
<tr>
<td>LR</td>
<td>Undersample</td>
<td>0.64 0.54 0.86</td>
<td>0.65 0.53 0.87</td>
</tr>
<tr>
<td>LR</td>
<td>No Balancing</td>
<td>0.57 0.7 0.52</td>
<td>0.64 0.68 0.61</td>
</tr>
<tr>
<td>MLP (10,1)</td>
<td>Oversample</td>
<td>0.65 0.54 0.86</td>
<td>0.64 0.53 0.85</td>
</tr>
<tr>
<td>MLP (10,1)</td>
<td>Undersample</td>
<td>0.64 0.53 0.85</td>
<td>0.67 0.54 0.85</td>
</tr>
<tr>
<td>MLP (10,1)</td>
<td>No Balancing</td>
<td>0.6 0.65 0.61</td>
<td>0.58 0.69 0.54</td>
</tr>
</tbody>
</table>
Conclusion and Future Scope
In conclusion, our results are consistent with the prior and current implementation of federated learning in healthcare domain\(^{2,3,4}\). Currently, our model architecture is limited to multilayer perceptron, whereas logistic regression is also a special case for that. We plan to extend our federated learning architecture to include more classifier algorithms such as support vector machines, decision trees, and random forest for future work. The major limitation of the analysis is the use of a relational database, which leads to under-representation of the challenges in the federated learning process in a real setting. Another key challenge is data inconsistencies, incompleteness, and lack of standardization. As future goals, we plan to use health systems that are already partitioned for better demonstration. Furthermore, we will explore other averaging techniques, optimal methods to address class-imbalance issues, better approaches for hyperparameter tuning of the global model, and extract other important patient-related factors from other data tables to continue our efforts building a more reliable and robust federated predictive model for diabetes-related complications. Our future work will also involve exploring and comparing other privacy-protected learning methods with federated learning.

References

<table>
<thead>
<tr>
<th>Model</th>
<th>Oversample</th>
<th>Undersample</th>
<th>No Balancing</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLP(10,10,1)</td>
<td>0.64</td>
<td>0.54</td>
<td>0.66</td>
</tr>
<tr>
<td>MLP(10,10,1)</td>
<td>0.62</td>
<td>0.55</td>
<td>0.69</td>
</tr>
<tr>
<td>MLP(10,10,1)</td>
<td>0.74</td>
<td>0.84</td>
<td>0.55</td>
</tr>
</tbody>
</table>


A Proposed Patient-Inclusive Methodology for Developing and Validating Telehealth Surveys that Include Social Determinants of Health

Izower, Mitchell, MD1,2, Liao, Zoe, B.S. Pharmacy1,3, Kim, Jeongeun, B.A.1,2, Quintana, Yuri, PhD1,2

1Beth Israel Deaconess Medical Center, Boston, MA, USA; 2Harvard Medical School, Boston, MA, USA; 3Northeastern University School of Pharmacy, Boston, MA, USA

Abstract

This paper discusses a method to develop and validate telehealth surveys that include social determinants of health domains. We performed a scoping review of literature on measuring social determinants of health and extracted 50 social determinants of health domains. We evaluated 14 validated telehealth surveys for questions associated with social determinants of health. We categorized the questions from the validated telehealth surveys using our extracted social determinants of health. We found that current validated telehealth-specific surveys only cover 16 (32%) of social determinants of health domains, with the most commonly evaluated domains being “Medical Needs” and “Social Connections/Isolation”. Telehealth services are a valuable modality to provide care to patients. Surveying patients is integral to performing quality improvement and improving patient outcomes. Social determinants of health are important factors in determining patient outcomes. We propose an approach to validating the missing domains and evaluating survey validity.

Introduction

The COVID-19 pandemic has greatly increased the use of telehealth services1. Telehealth services have several benefits over in-person healthcare delivery, such as reducing resource use at health centers, improving access to care, and minimizing the risk of disease transmission from person to person2. If the appropriate measures are taken in the realms of workforce development and support, consumer empowerment, healthcare funding reform, improvements to the digital health ecosystem, and telehealth integration into routine care, then the expanded use of telehealth will persist beyond the COVID-19 pandemic3.

We must regularly and comprehensively assess telehealth services to ensure that telehealth services continue to deliver high-quality care. Surveys can be used to measure patient satisfaction, experience, preference, and attitudes, and to assess the technical quality of the telehealth visit4. The major strengths of surveying include confidentiality, independent assessment of clinicians and patients, pre-existing scales which allow cross-study comparison, low cost relative to sample size, result generalizability, and the ability to validate surveys, thus ensuring accurate measurements of desired constructs5.

While surveys can identify valuable information for improving telehealth services, they also have limitations. Inappropriately constructed surveys may exhibit a lack of response variability5, poor reliability6, inapplicable response scale design5, lack of construct validity5, poor selection of constructs5, question ambiguity5, biased questions5, significant “halo” effect5, and other technical issues. Therefore, it is recommended that validated surveys be used to assess telehealth services5.

Validated surveys have been assessed and modified to ensure that their questions achieve face validity, usefulness, construct validity, convergent and discriminant validity, and content validity, and that the overall survey instrument achieves reliability and internal consistency5. Validated surveys have been developed to ensure that survey measurement of constructs is accurate, dependable, and reliable6,7. A recent systematic review identified 12 validated telehealth surveys8. Another review of telehealth research identified an additional 4 telehealth-specific surveys8. Examples of validated surveys include the “Patient Assessment of Communication during Telemedicine” (PACT)9 and “Telemedicine Satisfaction and Usefulness Questionnaire” (TSUQ)4.

However, while these validated telehealth-specific surveys provide valuable confidence in survey meaningfulness, they, too, still have shortcomings. Namely, while these surveys may evaluate patient experience and satisfaction, they may fail to identify underlying factors that contribute to those scores. One such group of underlying factors are social determinants of health.

The World Health Organization defines social determinants of health (SDoH) as “the conditions in which people are born, grow, live, work and age”. SDoH can contribute to health inequalities if they are not adequately addressed9.
is vitally important that telehealth services seek to evaluate the social determinants of health of patients using them. For this reason, it is important to consider if current validated telehealth-specific surveys can evaluate social determinants of health domains.

This paper evaluates current telehealth-specific validated surveys for coverage of social determinants of health. It seeks to provide recommendations for improving the evaluation of social determinants of health using validated telehealth-specific surveys. We propose a method for researchers to develop validated telehealth-specific surveys which include social determinants of health domains.

Methods

Our objective is to understand how current telemedicine surveys covered social determinants of health domains. To identify SDOH domains, we conducted a scoping literature search in PubMed with the search terms “social determinants of health,” and “screening,” “documentation,” “electronic health records,” or “data collection.” We reviewed papers where data was collected on at least one social determinant of health domain in a clinical setting as part of the care delivery process, and data were recorded in the electronic health record (EHR). Twelve studies covering 18 health organizations were included for analysis, and from these studies, 50 SDoH domains were identified.

We then searched the literature for validated telehealth-specific surveys. From this search, we identified 16 validated telehealth-specific surveys. One of the recommended surveys was about attitudes towards the Internet, and one was about the efficacy of email, so we excluded these two. The surveys were divided amongst each group member (4 people), and questions from each survey were correlated with a social determinant of health using a methodology derived from clinical and professional judgment. The final association between question and social determinant of health was reviewed by a physician, pharmacist, and Ph.D. with experience in medical informatics. Data analysis was then performed on the 14x50 matrix to see how the survey questions covered the SDoH previously identified. We subsequently developed a method to develop validated new survey questions for some of the missing domains.

Results

In reviewing the 14 validated telehealth surveys, we found that the questions only covered 16 (32%) of the 50 social determinants of health domains that we identified in Table 1. Of the covered domains, “Medical Needs” was covered by the most surveys (12, 86%), followed by Social Connections/Isolation (10, 71%), and Literacy/Learning Style (8, 57%).

Table 1. Surveys per social determinant of health domain.

<table>
<thead>
<tr>
<th>Domain</th>
<th>Surveys per Domain</th>
<th>Percent of Surveys (total studies 14)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medical needs (including health insurance)</td>
<td>12</td>
<td>86%</td>
</tr>
<tr>
<td>Social connections/isolation</td>
<td>10</td>
<td>71%</td>
</tr>
<tr>
<td>Literacy/Learning Style</td>
<td>8</td>
<td>57%</td>
</tr>
<tr>
<td>Health Literacy</td>
<td>6</td>
<td>43%</td>
</tr>
<tr>
<td>Transportation</td>
<td>5</td>
<td>36%</td>
</tr>
<tr>
<td>Discrimination</td>
<td>5</td>
<td>36%</td>
</tr>
<tr>
<td>Stress</td>
<td>4</td>
<td>29%</td>
</tr>
<tr>
<td>Utilities</td>
<td>4</td>
<td>29%</td>
</tr>
<tr>
<td>Hearing</td>
<td>3</td>
<td>21%</td>
</tr>
<tr>
<td>Vision</td>
<td>3</td>
<td>21%</td>
</tr>
<tr>
<td>Inadequate material resources</td>
<td>2</td>
<td>14%</td>
</tr>
<tr>
<td>Financial resource strain</td>
<td>1</td>
<td>7%</td>
</tr>
<tr>
<td>Residential address (neighborhood)</td>
<td>1</td>
<td>7%</td>
</tr>
<tr>
<td>Activities of daily living</td>
<td>1</td>
<td>7%</td>
</tr>
<tr>
<td>Caregiver stress</td>
<td>1</td>
<td>7%</td>
</tr>
<tr>
<td>Family stress</td>
<td>1</td>
<td>7%</td>
</tr>
</tbody>
</table>

The surveys that covered the most SDoH were the TESS\textsuperscript{10} and SUTAQ\textsuperscript{11} surveys, which covered 8 of 50 (16%) of SDoH domains each, followed by the TMPQ\textsuperscript{12} survey (7, 14%), as denoted in Table 2.
Table 2. Social determinants of health domains covered per survey.

<table>
<thead>
<tr>
<th>Survey</th>
<th>Covered SDoH Domains</th>
<th>Percent of domains (Domains = 50)</th>
<th>Abbreviated Domains</th>
</tr>
</thead>
<tbody>
<tr>
<td>TESS</td>
<td>8</td>
<td>16%</td>
<td>S.C., M.N., T, H.L., L.L., H., V., D.</td>
</tr>
<tr>
<td>SUTAQ</td>
<td>8</td>
<td>16%</td>
<td>S., M.N., T., U., H.L., A.D., C.S., D.</td>
</tr>
<tr>
<td>TSUQ</td>
<td>6</td>
<td>12%</td>
<td>S., S.C., M.N., U., L.L., D.</td>
</tr>
<tr>
<td>TUQ</td>
<td>6</td>
<td>12%</td>
<td>S.C., M.N., T., L.L., H., V.</td>
</tr>
<tr>
<td>TSQ</td>
<td>6</td>
<td>12%</td>
<td>S.C., M.N., T., L.L., H., V.</td>
</tr>
<tr>
<td>PACT</td>
<td>5</td>
<td>10%</td>
<td>S., S.C., M.N., T., H.L.</td>
</tr>
<tr>
<td>PEPPPI</td>
<td>4</td>
<td>8%</td>
<td>S.C., M.N., H.L., D.</td>
</tr>
<tr>
<td>TISQ</td>
<td>3</td>
<td>6%</td>
<td>S.C., M.N., H.L.</td>
</tr>
<tr>
<td>SUS</td>
<td>3</td>
<td>6%</td>
<td>S.C., M.N., L.L.</td>
</tr>
<tr>
<td>ACCP ICU</td>
<td>2</td>
<td>4%</td>
<td>R.A., I.R.</td>
</tr>
<tr>
<td>Telemedicine Survey</td>
<td>2</td>
<td>4%</td>
<td>M.N., L.L.</td>
</tr>
<tr>
<td>TAM-12</td>
<td>2</td>
<td>4%</td>
<td>M.N., L.L.</td>
</tr>
<tr>
<td>CSUQ</td>
<td>1</td>
<td>2%</td>
<td>L.L.</td>
</tr>
</tbody>
</table>


Discussion

Social determinants of health have been shown to contribute up to 40% towards determining health outcomes. Currently, no SDoH screening tool has been shown to be validated for all dimensions of validity. As telehealth becomes a common modality of health care delivery post-COVID-19, social determinants of health that influence patient experience and satisfaction should be evaluated to reduce existing health disparities and to improve telehealth services. In our review of validated telehealth-specific surveys, only 32% of surveys had questions related to the 50 social determinants of health domains, which is not enough for a comprehensive evaluation.

Since current validated telehealth-specific surveys do not adequately evaluate social determinants of health domains, we recommend creating new validated telehealth-specific questions which could be added to extant validated telehealth surveys that include social determinants of health. During development, survey developers should consider the social determinants of health domains that they desire to assess. Determining which SDoH domains should be included requires engagement with both providers and patients. Over 4 decades ago, Dr. Warner Slack once said, “The basis for our use of the computer in medicine is the thesis that the largest and least utilized provider of healthcare is the patient,” emphasizing that patient experience should be assessed through the lens of the patient in a respectful and meaningful way. Decades later, we are still trying to find meaningful ways to involve patients in their healthcare. Meaningfully engaged patients, providers, and relevant stakeholders must be part of the process from the very beginning. The success of telehealth depends not only on the technology, but the alignment and inclusion of all stakeholders. SDoH domains may be conceptualized by working with patients, as doing so can produce effective telehealth innovations.

Some pertinent SDoH domains that affect telehealth services include inadequate material resources, utilities, health literacy, language preference, and financial resource strain. Optimizing utility access, such as broadband internet, to participate in telehealth services can encourage the participation of underrepresented populations. Providing technical support to improve the health literacy of patients with difficulty understanding how telehealth works can lower the barrier to use telehealth. Another approach for increased inclusivity is to expand the service to non-English speakers, as language barriers and the availability of interpreter services prevent the utilization of telehealth.
by non-English speakers in low-income communities. However, such expansion in the service comes with costs: cost of the equipment, cost of hiring interpreters, cost of including more physicians in the service.

Once SDoH domains are chosen, questions may be selected from non-telehealth-specific surveys (validated if possible) or by modifying existing validated telehealth-survey questions. A recent systematic review found 21 screening tools that evaluated at least 2 social risk factors and were administered in a clinical setting. Reliability or validity testing was available for 8 of those survey instruments, with two of the instruments (Partners in Health Survey and Women's Health Questionnaire) reporting results from validity or reliability pilot testing. If no previously prepared questions are acceptable, new questions may be constructed. During this process, speaking to patients within the study population, for example by focus groups, can help identify candidate questions. Questions selected for inclusion in the survey should be evaluated for their relevance to the desired SDoH domain, bias, readability, responsiveness to change, clarity, consistency, and sensitivity to difficult topics. It is recommended that some questions be phrased positively and others negatively to diminish the “halo” effect. Also, recall questions should be avoided, as respondents may have limited ability to recall the past. The questions should offer logically distinct response options, including an option for non-response. Often, survey questions use a five-point Likert scale for scoring, when doing so, the scales should be scored in the same direction to make sense when summed. Once selected, modified, and new survey questions are prepared, they may be added to an extant survey, or combined into a new survey instrument. The modified or new survey instrument should then be validated.

The validation process should include two pilot studies and a final validation study. The first pilot study should aim to assess face and content validity, the usefulness of survey items, remove items that perform poorly, and ensure the survey assesses all relevant constructs. The second pilot study reevaluates the survey for the above qualities after editing, and refines wording and layout. Finally, the validation study should establish the construct, convergent, discriminant, and criterion-related validity of questions. The validation study also assesses the survey instrument for parallel forms reliability, test-retest reliability, internal consistency reliability, including calculating Cronbach’s alpha, and psychometric properties. Preferably, the validation study should be performed on a large population which is similar to the population planned to be evaluated by the validated survey.

Planning must be done for how to best administer the survey. Surveys may be administered as either a self-completion questionnaire or interview. During the pilot and validation studies, consideration should be given to sampling frame and strategy, sample size, administration methods, means to improve response rate, data management and analysis decision, and disclosures to participants. To understand the feasibility of survey implementation, consideration should be given to the number of times the survey needs to be administered, and maintaining respondent anonymity. In addition, a comprehensive method of data entry should be specified prior to data collection. Double data-entry is recommended to reduce error. Data storage should be secure and adhere to regulations, including patient privacy regulations.

One particular limitation of studies making use of and validating telehealth survey instruments is their frequently small sample size. Unfortunately, small sample sizes may be due to a significant amount of pilot and feasibility studies, and results in issues when performing meta-analysis and with study generalizability. Small sample size is one of the major obstacles in conducting randomized controlled trials and sample surveys to evaluate the impact of telemedicine programs. Small sample size may also limit generalizability and create bias in study results. Future studies should strive to reach an adequate sample size to ensure the power of the analysis and generalizability of results.

Our study did have some limitations. We limited our review to the SDoH domains that were collected by screening tools for documentation in electronic health records. More research is needed to draw evidence-based conclusions on which social determinants can affect telehealth uptake and subsequently health outcomes. However, these domains may not be fully comprehensive. For example, our 50 SDoH domains did not include broadband internet access, which was recently proposed as a critical SDoH in the age of telehealth. As our social determinants of health domains were not fully comprehensive, some questions may not have been able to be categorized, while others may have benefitted from an alternative categorization.

**Conclusion**

Social determinants of health (SDoH) have been shown to contribute up to 40% towards determining health outcomes. Despite this, no SDoH screening tool has been evaluated for all types of validity. Of the 14 validated telehealth surveys evaluated, the questions only covered 16 (32%) of the 50 social determinants of health domains that we identified. Telehealth services are a valuable modality to provide care to patients, but to improve them we
will need more comprehensive, validated telehealth surveys which include SDoH questions. In addition to being validated, these surveys should also include patients in their development. We have proposed a patient-inclusive method for developing and validating telehealth surveys that include social determinants of health. Future work includes involving patients in the co-design process to determine SDoH, and to develop validated telehealth surveys which include comprehensive social determinants of health domains.

References
20. Hu PJ, Chau PY, Sheng OR, Tam KY. Examining the technology acceptance model using physician
acceptance of telemedicine technology. Journal of management information systems. 1999 Sep 1;16(2):91-112.
Semantic Expansion of Clinician Generated Data Preferences for Automatic Patient Data Summarization

Ashutosh Jadhav, Ph.D, Tyler Baldwin, Ph.D, Joy Wu, MD MPH, Vandana Mukherjee, Ph.D, Tanveer Syeda-Mahmood, Ph.D
IBM Almaden Research Center, San Jose, CA

Abstract

Patient Electronic Health Records (EHRs) typically contain a substantial amount of data, which can lead to information overload for clinicians, especially in high-throughput fields like radiology. Thus, it would be beneficial to have a mechanism for summarizing the most clinically relevant patient information pertinent to the needs of clinicians. This study presents a novel approach for the curation of clinician EHR data preference information towards the ultimate goal of providing robust EHR summarization. Clinicians first provide a list of data items of interest across multiple EHR categories. Since this data is manually dictated, it has limited coverage and may not cover all the important terms relevant to a concept. To address this problem, we have developed a knowledge-driven semantic concept expansion approach by leveraging rich biomedical knowledge from the UMLS. The approach expands 1094 seed concepts to 22,325 concepts with 92.69% of the expanded concepts identified as relevant by clinicians.

1 Introduction

Patient Electronic Health Records (EHRs) usually contain very detailed information and are a source of a large amount of clinical data for a patient. EHRs document various aspects of the patients’ clinical information such as the reason for visit, problem list, labs and test results, allergies, medications, etc. They include information over multiple patient encounters and care provided by different healthcare professionals. Although the information is valuable, an abundance of patient information can lead to an information overload condition for clinicians. In order to find information pertinent to the current case, clinicians may go over several notes, labs, and reports spread across multiple visits and years. In such cases, identifying the most relevant information for clinical decision making can be difficult and time-consuming. Moreover, current EHR systems often do not present this tremendous amount of patient data in a way that supports the clinical workflow or cognitive reasoning, and immensely large records can negatively affect clinical work due to error of omission or delay. Thus, it would be beneficial to have a mechanism for summarizing the most clinically relevant patient information. Previous studies have shown that EHR summaries can have a positive impact on overall patient care.

Selecting the relevant information from a patient record is a difficult problem. In this work, we propose a method for the curation and expansion of EHR data preferences, towards summarization based on clinician identified concepts. In the first step, clinicians manually generate a summarization blueprint, or a “summary template”, by specifying which information they would like to see in a holistic summary of a patient’s EHR. This patient information is captured by identifying important clinical concepts and their categories that are useful for generating summary documents. Since the summary template is generated manually, it has limited coverage and may not cover all the important terms relevant to a concept. For example, for the “diabetes” concept, there are multiple relevant concepts such as diabetes mellitus, high blood sugar, diabetes mellitus infantile, insulin pump, diabetes mellitus insulin dependent, insulin resistance, pregnancy induced diabetes, etc. Moreover, for a given seed concept, the related concepts may span across multiple categories such as medication, labs, and allergy.

The objective of this work is to semantically expand the clinician-generated initial summarization template in order to have broader coverage of clinically relevant concepts and categorize these concepts in the selected 11 clinical categories. Although powerful tools such as the Universal Medical Language System (UMLS) Metathesaurus are able to aid in the expansion of medical concepts, there is no universally agreed upon method for performing this expansion. The most straightforward expansion methods all have notable limitations. Keyword-based search on the UMLS Metathesaurus does not consider the semantics of the query and returns all the concepts containing the keyword; it does not rank concepts by considering the relevance of the expanded concepts to the seed concepts. And
while UMLS has relationships between nodes, these relationships are often vague and inconsistent in their granularity, making simple graph traversal perilous. Using UMLS semantic types and relationships provides a large number of possible related concepts that has to be filtered-out vigilantly to select relevant expansions. Thus, most naïve methods of semantic expansion would require a large amount of manual correction by clinician experts.

In order to address this problem, we propose the following multi-step summary template generation approach. First, clinician experts provide a list of important concepts of interest across multiple summarization categories. As these experts may not have knowledge of the UMLS structure, these concepts are initially provided in plain-text format rather than mapped to UMLS identifiers. In the second step, we use a clinical concept extractor to map these plain text terms to UMLS concept identifiers automatically. This automated mapping is then validated and corrected as necessary by the original clinical experts. These validated seed concepts are then automatically expanded by leveraging rich information from a UMLS-backed biomedical knowledge graph. The seed concepts are expanded by identifying their clinical variants and related concepts based on the hierarchy and relationships from the biomedical knowledge graph. Further, the expanded concepts are filtered by removing duplicate concepts and semantically irrelevant concepts. In the fourth step, the filtered concepts are categorized into selected clinical categories.

This work aims to utilize semi-automated clinician-in-the-loop systems in order to efficiently understand data preferences. By doing so, we are able to keep the burden on the clinical experts low while greatly expanding upon their seed concepts and ensuring that their initial intent is understood and met. Further, linking every term to one or more concepts in the knowledge base ensures that they can be mapped to EHR data from disparate sources. The proposed method is able to expand 1,094 seed concepts from 11 clinical categories to 22,325 expanded concepts. Also, this is done while keeping the relevancy rate at 92.69% and with 84.52% concepts categorized into correct clinical categories, which helps to reduce the validation burden. By streamlining data preference gathering and interpretation, summarization systems may more easily be tailored towards the needs of specific specialists. To the best of our knowledge, this is the first EHR data prioritization approach, which keeps the clinicians’ inputs as the focal point while leveraging the rich knowledge from a biomedical knowledge base. The data prioritization and expansion techniques described herein should be applicable across disciplines.

1.1 Related work

1.1.1 Patient summarization report generation

Summarization methods can be broadly categorized as extractive or abstractive. Extractive summaries are created by borrowing phrases or sentences from the original input text. In the domain of clinical summarization, an extractive approach can identify pieces of the patient’s record and display them without providing additional layers of abstraction. Abstractive summaries generate new text that synthesizes the original text. In the domain of clinical summarization, abstractive summaries may provide additional higher-level context to explain the data, such as computed quantities (e.g., trends) or automatically generated text. Much of the current research on summarization in the biomedical domain has focused on text summarization, in which one or more texts are reduced to a single condensed reference text. Text summarization strategies have been developed for automated summarization of scientific literature, generation of literature abstracts, and summarization and translation.

Another dichotomy in summary generation techniques is between methods based on knowledge and those based on data. The data-driven approaches require less context-specific knowledge to create the summary while “knowledge rich” methods necessitate larger and more advanced knowledge bases. A review by Mishra et al. indicated that there is a growing interest in knowledge-rich approaches in the biomedical domain, coinciding with the increased availability of comprehensive lexical resources, such as WordNet and the Unified Medical Language System (UMLS). Common techniques of extraction-based summarization include topic-based sentence extraction, where the relevance of a sentence is computed with respect to one or more topics of interest; Van Vleck et al. performed structured interviews to identify and classify phrases that clinicians considered relevant to explaining a patient's history.
1.1.2 Automatic expansion of medical terms:

In this paper, we present an approach to data expansion reliant on automatic semantic expansion of seed terms. This is similar to existing methods in information retrieval (IR) that falls broadly into the category of automatic query expansion (AQE)\(^6\). The majority of AQE techniques are tailored towards Web search and rely on relevancy feedback\(^7\), click-through data\(^8,9\) or similar queries\(^6\). Until recently most techniques used in AQE were “knowledge-poor”, but with recent advancements in knowledge graphs\(^20,21\) some AQE techniques have been developed to leverage this knowledge\(^22\).

In the medical domain, researchers have developed AQE techniques based on domain-specific ontologies and using the UMLS Metathesaurus. Most of the recent medical IR research has focused on developing knowledge-based (or concept-based) retrieval models dependent on medical resources such as the UMLS. Aronson and Rindflesch\(^23\) use the MetaMap program for associating UMLS Metathesaurus concepts with the original query. They conclude that the optimal strategy would be to combine query expansion with retrieval feedback. Many studies have also utilized the Medical Subject Headings (MeSH) thesaurus in query expansion\(^24\). Zhou et al.\(^25\) expanded query terms utilizing several vocabulary sources, including MeSH. Sondhi et al.\(^26\) used both the MeSH medical thesaurus and manual physician feedback in query expansion, comparing different combinations of methodology.

2 Methods

This section gives an overview of the proposed data preference gathering method based on summary template generation and expansion. Broadly, the major steps are as follows: 1) Generation of the seed summarization template, 2) Semantic expansion of the summarization template, and 3) Patient summary generation based on the expanded template.

<table>
<thead>
<tr>
<th>No</th>
<th>Clinical Category</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Allergy</td>
<td>Allergy to medication, contrast agents, food or other allergens</td>
</tr>
<tr>
<td>2</td>
<td>Family Member</td>
<td>List of “close”, usually first degree, family member, whose history of a heritable medical illness may be an indication of increased risk for the same illness for the patient</td>
</tr>
<tr>
<td>3</td>
<td>Family History</td>
<td>Heritable medical illnesses</td>
</tr>
<tr>
<td>4</td>
<td>Imaging</td>
<td>Radiology imaging related concepts</td>
</tr>
<tr>
<td>5</td>
<td>Implanted Devices</td>
<td>Medical devices implanted in the patient's body, such as a pacemaker</td>
</tr>
<tr>
<td>6</td>
<td>Labs</td>
<td>All non-imaging investigations such as blood and urine tests</td>
</tr>
<tr>
<td>7</td>
<td>Medications</td>
<td>The medications that the patient has been prescribed</td>
</tr>
<tr>
<td>8</td>
<td>Patient Management</td>
<td>Treatments and procedures (such as surgery)</td>
</tr>
<tr>
<td>9</td>
<td>Problem list</td>
<td>Patient's medical illness, signs and symptoms</td>
</tr>
<tr>
<td>10</td>
<td>Social History</td>
<td>Patient's social factors that may influence health outcome, such as smoking status, occupation and other at-risk behaviors</td>
</tr>
<tr>
<td>11</td>
<td>Vitals</td>
<td>Key physical exam parameters, such as heart rate, temperature and respiratory rate</td>
</tr>
</tbody>
</table>

Table 1. Clinical categories selected for the summarization template

2.1 Generation of seed summarization template

The seed summarization template is a minimal version of one or more clinical experts’ data preferences seeded by manual curation. To arrive at this initial template, one must first identify the clinical categories of interest and the terms within those categories that are most relevant to the clinical experts. The following are the major components to the generation of the seed summarization template:

- Identification of Clinical categories: The clinical categories of interest may vary between specialties. In each iteration of the data preference gathering technique described herein, the clinical experts provide feedback to help select the appropriate categories. In any examination of electronic health records, the clinical variables of interest can be described in terms of a set of known clinical categories denoted as \(\{C_1, C_2, \ldots, C_n\}\). The clinical categories used in this work are shown in Table 1. These categories were first seeded by those commonly used in the history
and physical examination and were then modified based on input from radiologists who participated in the seed summary template generation.

- Term identification: During the term identification stage, 10 practicing radiologists worked collaboratively to identify data items in each category that they felt were important to see as part of a clinical summary, focusing specifically on what data was of interest in their discipline. These terms were provided as plain text explanations with varying levels of specificity. For instance, categories such as Physical Exam Findings could contain both terms specific to a particular finding such as murmurs, or a much more general statement of interest such as enlarged solid organs.

- Term to clinical concept (CUI) mapping: Terms identified by clinicians are free text English terms and in order to understand their semantics we need to map these terms to their respective clinical concepts in a biomedical knowledge base. Using a medical concept extractor, we mapped the clinician identified terms to UMLS concepts.

- Concept mapping review: With automatic term to CUI mapping, each term can be mapped to multiple concepts or may fail to be mapped as at all. Therefore, to confirm that the terms are mapped to the correct concepts, the mapping is reviewed by the clinicians who first generated the seed terms. A summarization template \( T \) is then defined as collection of clinical concepts linked seed concepts corresponding to their clinical category \( C_i \).

2.2 Semantic expansion of the summarization template

A multi-fold, automated semantic term expansion methodology is employed in order to improve the coverage of the seed summarization template. The following are the major components of the semantic template expansion approach: 1) Biomedical knowledge graph, 2) Clinical variant generation, 3) Hierarchical expansion, 4) Concept filtration, 5) Concept categorization, and 6) Review and validation of expanded template concepts

2.2.1 Biomedical knowledge graph.

The core of our biomedical knowledge graph is the UMLS\(^6\). The National Library of Medicine (NLM) produces the UMLS to facilitate computer understanding of biomedical text. The UMLS is a repository of more than 100 biomedical vocabularies. The UMLS consists of the following three subcomponents: the metathesaurus, the semantic network, and the SPECIALIST lexicon. The Metathesaurus forms the base of the UMLS and comprises of over 4 million biomedical concepts and 16 million concept names, all of which stem from the over 130 incorporated controlled vocabularies and classification systems. The Semantic Network consists of semantic types and semantic relationships. Semantic types are broad subject categories like “Disease or Syndrome” and “Clinical Drug”. Semantic relationships are useful relationships that exist between Semantic Types. Each concept in the Metathesaurus is assigned one or more semantic types (categories), which are linked with one another through semantic relationships. The SPECIALIST lexicon is an English-language lexicon that contains biomedical terms. The lexicon entry for each word or term records the syntactic, morphological, and orthographic information of the respective lemma. It also contains spelling variants, acronyms, and abbreviations.

The biomedical knowledge base used in this work contains a subset of the UMLS, generated as follows: all vocabularies with license type 0 as defined by UMLS are selected along with SNOMED CT. All vocabularies in this set that are not in English are then removed. Our biomedical knowledge base also includes Radlex\(^7\), a radiology-specific vocabulary not typically contained within UMLS. To link terms from this vocabulary to the UMLS-based concepts in the rest of the knowledge base, we utilize the linking provided by the National Cancer Institute Metathesaurus (NCIm)\(^8\).

2.2.2 Clinical variant generation

In the initial summarization template generation step, all the terms identified by clinicians are mapped to clinical concepts in the knowledge base. In the variant generation process, clinical variants of each seed concept are generated by identifying synonymous concepts. In the UMLS, each concept is associated with a single unique identification string referred as its Concept Unique Identifier (CUI). This CUI may be linked to multiple concepts from the source vocabularies that feed into UMLS, which may be described by different term variants. For instance, UMLS CUI
C0020538 “Hypertensive disease” is linked to multiple term variants such as “high BP”, “high blood pressure”, and “hypertension”. The clinical variant generation collects all of these synonymous concepts.

In the second phase of clinical variant generation, related concepts are identified. The candidate-related concepts are assembled by a keyword-based search performed on the knowledge-base using the lexical variants of the seed concepts and their clinical concept variants as inputs. The keyword-based search on the clinical knowledge-base returns numerous concepts. In order to maintain the relevancy of the concepts, from the search results, we select all the concepts that encompass either the seed terms or their clinical variants (e.g., for “diabetes”, get all the concepts that contains “diabetes”) and select only top 20 concepts from the partial match results. While this particular step helps us to improve our recall for the concept expansion to identify relevant concepts, it also identifies a lot of irrelevant concepts. For example, for “diabetes”, this step identifies concepts like “diabetes insipidus” and “vasopressin resistant diabetes insipidus” which are clinically not relevant to “diabetes”. Therefore, the candidate-related concepts are filtered out based on the semantic and textual features as described in the “Template concept filtration” section.

2.2.3 Hierarchical expansion

Hierarchical expansion involves identification of parent, sibling and children concepts for the clinical variants. Concepts in the UMLS are connected to each other using semantic relationships in the semantic network. For each concept, we can retrieve the concept’s parent and children concepts by traversing the concept hierarchy of a source vocabulary (such as SNOMED CT). Using the seed’s parent concept, we further identify the siblings of the seed term. The concept hierarchy traversal is done for two hops and parent/children concepts are retained only if their semantic type matches with the starting concept. This helps to keep the scope narrow and to get relevant concepts. For all the concepts retained in the concept hierarchy traversal, we further identify their clinical variants as described previously. Figure 1 illustrates an extract of the relevancy graph generated for the seed term “diabetes” that maps to the concept ‘diabetes mellitus’ in the knowledge graph. As shown, the sphere of influence for a semantic expansion includes synonyms and many similar terms that may have been intended to be captured by the original clinician-provided phrase.

![Figure 1. Illustration of semantic expansion of a seed term (“diabetes”).](image-url)
2.2.4 Concept filtration

During the clinical variant generation and hierarchical expansion steps multiple synonymous and related concepts are identified as relevant for the seed concepts. In this step, duplicate and semantically irrelevant concepts are filtered out to reduce noise. First, all concepts are converted to singular form (e.g., medications to medication). Duplicate concepts are filtered out based on the syntactic and morphological information of the concepts. If multiple concepts contain the same set of words in a different order (e.g., “high blood pressure”, “blood pressure, high”), all but one of these concepts is filtered out.

Semantically irrelevant concepts are identified by examining the semantic information associated with each concept in the knowledge base as well as some empirically discovered textual features. The semantic features employed include the semantic type of the concept, the distance in hops from the concept to the original seed concept, and the UMLS relationship types between the concepts. We have selected a list of semantic types from UMLS that are relevant to EHR summarization such as “Clinical Drug”, “Laboratory or Test Result”, “Diagnostic Procedure”, “Disease or Syndrome”, and “Medical Device”. We retain only those concepts that have semantic types from our selected list of semantic types. Second, we compute the distance from the expanded concept to the seed clinical concept in the knowledge graph and we retain only those concepts that are less than or equal to 3 hops. This step helps us to filter out “diabetes insipidus” as a related concept for “diabetes” since the distance between these concepts in the knowledge-graph is not less than or equal to 3 hops. Finally, we perform filtration based on selected UMLS relationship types.

For textual features, several stop words indicative of semantic drift were identified by experimentation. Examples of identified stop words include “institute”, “doctor”, “nurse”, “admission”, “hospital”, “education”, and “facility”. The semantic and textual features were selected and tuned based on iterative empirical analysis of relevant and non-relevant concepts.

2.2.5 Concept categorization

In the concept categorization step, the concepts are categorized into the selected 12 clinical categories. This work is inspired by previous approach that uses UMLS semantic types for the concept categorization\(^2^9\). As mentioned earlier, we have curated a list of semantic types that relevant for EHR summarization. We mapped these selected semantic types to one or more clinical categories from (Table 1). We found direct mapping to one or more of from the selected list of semantic types for most of the health categories e.g., for the “Medication” clinical category, we have mapped “Pharmacologic Substance”, and “Clinical Drug” semantic types. For some of the categories, we have used a combination of semantic types and concepts for the categorization, e.g., for “Social history” category, we have used semantic types like “Occupational Activity” and concepts/keywords like “smoking”, “substance abuse”, and “illicit drug”.

2.2.5 Review and validation of expanded template concepts.

Once potentially relevant concepts have been identified via the above expansion process, they are given to clinician annotators for review. Two clinicians were involved in the review process. To reduce the probability of human errors and subjectivity, the two clinicians worked together to set-up an annotation scheme and annotated the first 250 concepts collaboratively. First, the clinicians labeled each term as relevant or not relevant for the template. Second, for each relevant concept, the clinicians checked the assigned clinical category of the concept based the category description and label it correct/incorrect. The clinicians then worked independently to annotate the rest of the expanded concepts from the dataset. Each concept was annotated by two clinicians. After all the annotations were completed, the annotators discussed and resolved conflicts on the concepts that had mismatching labels. This step is further elaborated on in the Results section.

2.3 Use-case - Patient summarization report generation

Once the expanded template is complete, it can be used in conjunction with EHR extraction tools to generate a patient summarization report. If a piece of information in the patient record is linked to a UMLS CUI, the expanded summarization template can be used as a guide to understand its relevancy. All the assembled concepts from the patient’s electronic health record can be matched against the concepts from the expanded template, with only matched
concepts retained in the summary. Alternatively, the extracted information could be ranked based on template-guided factors, such as whether the patient information matches concepts specified by the clinician in the initial template or the expanded concepts or the depth of the knowledge graph to which the concept matches. Since much of the clinical information is buried in free text such as in surgical notes or discharge summaries, having a list of important concepts would be very valuable. Furthermore, since the concepts are organized by the clinical categories, one approach would be to extract relevant sections of EHR documents and use concepts from related clinical categories for the summarization purpose.

3 Results

The seed template generated based on methodology described in section 2.1 had 1094 concepts. The distribution of these seed concepts across the selected clinical categories is listed in the 2nd column of the (Table 2). After all the steps in the template expansion process were completed, the 1,094 concepts in the seed template were expanded to a total of 22,325 concepts. The (Table 2), provides a summary results of our template expansion work. The “number of relevant concepts” is computed based on the concept annotations (section 2.2.5) while the “percentage of relevant concepts” is computed as \( \frac{\text{number of relevant concepts} \times 100}{\text{number of concepts after expansion}} \). The “number of relevant and correctly categorized concepts” indicates the number of concepts that clinician found 1) relevant for the summarization template and 2) categorized into correct clinical categories. The “percentage of relevant and correctly categorized concepts” is computed as \( \frac{\text{number of relevant and correctly categorized concepts} \times 100}{\text{number of concepts after expansion}} \). Finally, the “average number of relevant and correctly categorized concepts per seed” is computed as \( \frac{\text{number of relevant and correctly categorized concepts}}{\text{number of seed concepts}} \).

<table>
<thead>
<tr>
<th>Clinical categories</th>
<th>Number of seed concepts</th>
<th>Number of concepts after expansion</th>
<th>Number of relevant concepts</th>
<th>Percentage of relevant concepts</th>
<th>Number of relevant and correctly categorized concepts</th>
<th>Percentage of relevant and correctly categorized concepts</th>
<th>Average number of relevant and correctly categorized concepts per seed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allergy</td>
<td>16</td>
<td>129</td>
<td>125</td>
<td>96.90%</td>
<td>113</td>
<td>87.60%</td>
<td>7.06</td>
</tr>
<tr>
<td>Family History</td>
<td>66</td>
<td>385</td>
<td>313</td>
<td>81.30%</td>
<td>313</td>
<td>81.30%</td>
<td>4.74</td>
</tr>
<tr>
<td>Family Member</td>
<td>21</td>
<td>107</td>
<td>80</td>
<td>74.77%</td>
<td>80</td>
<td>74.77%</td>
<td>3.81</td>
</tr>
<tr>
<td>Imaging</td>
<td>119</td>
<td>1196</td>
<td>1038</td>
<td>86.79%</td>
<td>948</td>
<td>79.26%</td>
<td>7.97</td>
</tr>
<tr>
<td>Implant Devices</td>
<td>86</td>
<td>1452</td>
<td>1376</td>
<td>94.77%</td>
<td>1308</td>
<td>90.08%</td>
<td>15.21</td>
</tr>
<tr>
<td>Labs</td>
<td>83</td>
<td>1396</td>
<td>1278</td>
<td>91.55%</td>
<td>1190</td>
<td>85.24%</td>
<td>14.34</td>
</tr>
<tr>
<td>Medication</td>
<td>66</td>
<td>1058</td>
<td>1028</td>
<td>97.16%</td>
<td>936</td>
<td>88.47%</td>
<td>14.18</td>
</tr>
<tr>
<td>Patient Management</td>
<td>270</td>
<td>5160</td>
<td>4768</td>
<td>92.40%</td>
<td>4203</td>
<td>81.45%</td>
<td>15.57</td>
</tr>
<tr>
<td>Problem list</td>
<td>332</td>
<td>10468</td>
<td>9746</td>
<td>93.10%</td>
<td>8939</td>
<td>85.39%</td>
<td>26.92</td>
</tr>
<tr>
<td>Social History</td>
<td>21</td>
<td>516</td>
<td>497</td>
<td>96.32%</td>
<td>442</td>
<td>85.66%</td>
<td>21.05</td>
</tr>
<tr>
<td>Vitals</td>
<td>14</td>
<td>458</td>
<td>445</td>
<td>97.16%</td>
<td>397</td>
<td>86.68%</td>
<td>28.36</td>
</tr>
<tr>
<td>Total</td>
<td>1094</td>
<td>22325</td>
<td>20694</td>
<td>92.69%</td>
<td>18869</td>
<td>84.52%</td>
<td>17.25</td>
</tr>
</tbody>
</table>

Table 2. Summary result for template concept expansion by clinical categories

As shown in (Table 2), the number concepts, their expansion, relevancy percentage, and categorization performance differ with the clinical categories. The “Problem list” and “Patient Management” categories had most number of seed concepts while “Allergy”, “Family Member”, “Social history”, and “Vitals” categories had least number of seed concepts. The “percentage of relevant concepts” is over 90% for most the categories, while “Family history”, and “Family member” had lowest “percentage of relevant concepts”. The outcome of “percentage of relevant and correctly categorized concepts” is over 90% for most the categories.
categorized concepts” is dependent on the performance of each of the previous modules (clinical variant generation, hierarchical expansion and concept filtration) as well as our categorization approach. We found that 84.52% of the expanded concepts are relevant for the summarization template and categorized into correct clinical categories. The “average number of relevant and correctly categorized concepts per seed” shows the average concept expansion rate for relevant and correctly categorized concepts by clinical categories. The “Problem list” and “Vital” categories had most average concept expansion rate (more than 26 concepts per seed concepts) while “Family history” and “Family member” categories had least average concept expansion rate (less than 5 concepts per seed concepts). During the validation step, clinicians manually corrected the categories of misclassified concepts and removed incorrect concepts resulting in a final template of 20,694 validated concepts.

Although the validation of the expanded concepts gives some indication of the overall precision of the method, it is unable to assess its recall. A full understanding of recall would need ground truth information about all concepts that are relevant to each seed concept, which would be a taxing and likely unreliable thing to produce via manual clinician input. However, in order to get a rough understanding of recall, we performed a manual analysis on a small scale. We randomly selected twenty seed terms from the original summarization template, and two clinicians manually expanded these selected terms. Since manual expansion between two clinicians for the selected few terms did not match (varied by multiple terms), we could not calculate the recall statistics. However, we discovered that our automated approach covered, on average, 94% of the terms mentioned by the clinicians and many relevant terms that clinicians missed. The 6% terms that were missed in the automated expansion were related to the seed terms but were separated by multiple hops in the knowledge graph.

4 Discussion and Conclusion

Patient Electronic Health Records (EHRs) usually contain very detailed information. Although the information is valuable, an abundance of patient information can lead to an information overload condition for clinicians. Thus, it would be beneficial to have a mechanism for summarizing the most clinically relevant patient information pertinent to the needs of the clinicians. For example, let us consider a real-world use-case with radiologists. Radiologists today rely mostly on their visual interpretation of imaging studies for their diagnostic decisions. There is considerable information about the patient in the electronic health record that could positively impact their decisions. However, this data is often physically distributed across many enterprise hospital systems including electronic medical record (EMR) systems, radiology informatics systems (RIS), picture archiving and communication systems (PACS), laboratory systems, pharmacy systems, etc. Due to the volume of imagery to examine, radiologists have little time themselves to assimilate all pertinent clinical information from these distributed records. As a result, overlooking of diagnosis and misinterpretation is a common problem contributing to diagnosis error rates. This problem is further exacerbated by the differing needs among clinicians. Much of the information of note to a primary care physician would not impact a radiologist’s diagnosis, but certain domain-specific pieces of knowledge, such as an allergy to imaging contrast, could be critically important. If a radiologist’s visual examination could be augmented by providing a compact summary of the patient’s clinical history that focused on their specific data needs, it could lead to improvement in overall clinical decision making.

To this end, this work presents a hybrid approach that leverages biomedical knowledge and expert (clinician) knowledge for efficiently gathering clinical data preferences for use in the production of a EHR data summary. To do so, we introduce the notion of a “summarization template”, a catalogue of data preferences linked to a knowledge base. Summarization template is crafted semi-automatically with clinician assistance. An initial summarization template is manually generated by clinicians and is a high-level specification of patient information that the clinician would like to see in a holistic summary of the patient’s EHR. In its initial state, the template is incomplete and specifies the expected information in free text without linking to a knowledge base. If used as such, it would have limited coverage within an actual patient record. To alleviate this, seed concepts from the initial template are automatically linked to the knowledge base, validated by clinicians, and further expanded with related concepts. The expanded template is then reviewed and validated by clinical experts before it is considered to be ready for use in EHR summarization.

Limitations and outlook: In this work, we have not presented evaluation of the summaries generated using the summarization template due to scope of this work. In section 2.3 we have discussed an approach for EHR summarization using the summarization template. In future, we are planning to report evaluation of the summaries
generated using the summarization template and investigate if specialty templates (e.g., cardiology, neurology) are more helpful in summarization than a generic summarization template.

This study provides a novel approach to EHR data preference selection, keeping the clinicians at the center. It presents a knowledge-driven semantic expansion approach that leverages a rich biomedical knowledge base to expand upon manually identified seed terms. The approach expands 1,094 seed concepts to 20,694 clinically relevant concepts. Results of a validation study indicate that the expansion technique is able to identify relevant concepts with a relevancy rate of 92.69% and able to correctly categorize these concepts in 84.52% of cases, keeping the burden of the clinical validators fairly low. By iteratively relying on clinical expertise and automatic semantic expansion, the proposed method is able to generate and validate specialty-specific data preferences, which can then be used to tailor the output of EHR summarization systems to the needs of each specialist. The approach presented in this work can be used to curate custom vocabularies for different specialties of medicine like cardiology vocabulary, neurology vocabulary. Another possible application of this work would be in improving EHR document annotation and understanding. The approach presented in work is transferable to (and replicable) to address different problems in healthcare field that may benefit from domain/problem-specific custom vocabularies.

References
Learning Predictive and Interpretable Timeseries Summaries from ICU Data

Nari Johnson, MSc, Sonali Parbhoo, PhD, Andrew S Ross, PhD, Finale Doshi-Velez, PhD
School of Engineering and Applied Sciences, Harvard University, Cambridge, MA

Abstract

Machine learning models that utilize patient data across time (rather than just the most recent measurements) have increased performance for many risk stratification tasks in the intensive care unit. However, many of these models and their learned representations are complex and therefore difficult for clinicians to interpret, creating challenges for validation. Our work proposes a new procedure to learn summaries of clinical timeseries that are both predictive and easily understood by humans. Specifically, our summaries consist of simple and intuitive functions of clinical data (e.g., “falling mean arterial pressure”). Our learned summaries outperform traditional interpretable model classes and achieve performance comparable to state-of-the-art deep learning models on an in-hospital mortality classification task.

1 Introduction

Accurate predictions of patient risk in critical care units can aid clinicians in making more effective decisions. Specifically, early identification of patients at high risk for in-hospital mortality is critical to assess patient disease acuity and inform life-saving interventions [1, 2]. To predict in-hospital mortality risk, researchers have developed algorithms ranging from simple score-cards [1, 3] to statistical machine learning (ML) models. Recent advances in ML have led to the development of models with vast improvements in predictive accuracy for patient in-hospital mortality risk [4–7].

Despite these improvements, however, ML models are still prone to critical errors, often failing to generalize across different care settings or institutions [8]. An emerging line of research in interpretability, defined by [9] as “the ability to explain or to present in understandable terms to a human,” provides an alternative way to ensure systems preserve properties such as safety or nondiscrimination. If a model is interpretable to stakeholders, then clinical experts can inspect the model and verify that its reasoning is sound. This ability to audit and validate is especially important when the models are used to inform critical decisions affecting patient health.

In this work, we present a novel ML method to learn clinical timeseries summaries that are interpretable and predictive. We introduce functions to compute summaries that align with simple and intuitive concepts, such as whether the timeseries is decreasing or spikes above a critical threshold. In contrast to prior work, our method learns how much of the timeseries should be used to calculate these summaries, discarding earlier timesteps that may be irrelevant for a specific prediction task. Importantly, we introduce relaxations of our summary definitions to enable differentiable optimization, allowing summary parameters to be learned jointly with those of a downstream model. We show that with our method, we can achieve accuracies comparable with state-of-the-art baselines without sacrificing interpretability.

2 Related Work

Prior work on explaining clinical timeseries models fall into two categories: learning simple models that are inherently interpretable, and generating explanations of complex black box models. We summarize a few key examples below.

Interpreting Deep Models. One popular strategy for explaining ML models is learning a second post-hoc explanation model to explain the first black box model [10]. Many post-hoc explanation techniques for clinical timeseries models train explanation models that quantify the relative importance of each clinical variable [11, 12]. However, several works argue against the use of post-hoc explanation techniques, as explanation models are not always faithful or representative of the true underlying black boxes [13, 14]. Our method avoids these problems by design, instead explicitly optimizing for interpretability so that a second explanation model is not needed.

Another line of research proposes attention mechanism models specifically designed for timeseries. [15, 16] present neural attention architectures for clinical timeseries and argue that attention scores measure feature importance. However, attention methods are often highly complex and nonlinear. Furthermore, [17] shows attention scores do not always reflect true importance. Instead of approximating importance, our study uses linear models over richer fea-
tures, where importance does not need to be approximated but can be directly read off model coefficients.

**Expert Systems and Expert Features.** Our work extends a long tradition of clinical experts hand-crafting features to create interpretable clinical decision-support algorithms. Two expert systems widely used in ICUs are SAPS-II [1] and APACHE [3], which use simple score-card algorithms to evaluate patient acuity. These systems use input features such as the patient’s average or worst lab or vital values over time to compute mortality risk. While SAPS-II and APACHE are simple and simulable to clinicians, their predictiveness is limited, as they cannot capture how labs or vitals change across a patient’s stay.

A similar line of research proposes manual construction of expert features from clinical data, which are then used as input to ML models [18, 19]. One limitation of this approach is that expert feature derivation is expensive and requires clinical expertise. Rather than relying on expert knowledge to identify which summary features will be the most predictive for a given task, our work instead uses optimization with a sparsity constraint to automatically learn which summary features are the most predictive.

**Summarizing Clinical Timeseries.** A number of works have proposed a wide range of summary statistics for patient timeseries data. Many works such as [4, 20] train clinical models using the minimum value, maximum value, first measured value, or skew of clinical timeseries data. [21] proposes a more comprehensive set of 14 summary statistics to characterize the central tendency, dispersion tendency, and distribution shape of clinical timeseries data. Our work extends this research, and is the first to our knowledge to use the slope of the timeseries or proportion of time above or below critical thresholds. Our work is also novel in that we explicitly model and optimize for the duration over which we compute each summary feature.

### 3 Cohort and Problem Set-Up

Our goal is to learn summaries of patient timeseries data that are both human-interpretable and predictive for a downstream classification task. Our approach will define how to calculate these human-interpretable summaries, and describe how both summary and classification model parameters are learned through optimization. In what follows, we detail each of these processes.

**Prediction Task.** Our work examines early prediction of in-hospital mortality. We use the patient’s first 24 hours of data to predict if they would later expire over the course of the remainder of their admission. Patients who expired in the first 24 hours of their stay were excluded from our cohort.

**Cohort Selection.** We use data from the MIMIC-III critical care database [22], which contains deidentified health data from patients in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012. All data was extracted from MIMIC-III PhysioNet version 1.4, which contains 30,232 patients. We exclude patients under 18 years of age and patients whose weight is not measured. We include data from each patient’s first hospitalization only, and only patients with stays between 24-72 hours [4]. After applying these criteria, our final cohort contained 11,035 patients, 15.23% of whom died in-hospital. Cohort characteristics and demographics are summarized in Table 1.

<table>
<thead>
<tr>
<th>Cohort</th>
<th>Age</th>
<th>% Female</th>
<th>% Urgent</th>
<th>% Emergency</th>
<th>% Elective</th>
<th>% MICU</th>
<th>% SICU</th>
<th>% CCU</th>
<th>% CSRU</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>64.7</td>
<td>43.8</td>
<td>1.12</td>
<td>84.46</td>
<td>14.43</td>
<td>42</td>
<td>18</td>
<td>12</td>
<td>16</td>
</tr>
<tr>
<td>+</td>
<td>70.9</td>
<td>46.5</td>
<td>0.77</td>
<td>96.25</td>
<td>2.97</td>
<td>53</td>
<td>18</td>
<td>13</td>
<td>4</td>
</tr>
<tr>
<td>-</td>
<td>64.1</td>
<td>43.6</td>
<td>1.14</td>
<td>83.22</td>
<td>15.64</td>
<td>41</td>
<td>18</td>
<td>12</td>
<td>17</td>
</tr>
</tbody>
</table>

**Table 1:** Mean statistics for the population cohort, and for cohorts of positive versus negative patients for in-hospital mortality. Abbreviations: MICU, medical care unit; SICU, surgical care unit; CCU, cardiac care unit; CSRU, cardiac-surgery recovery unit.

### 3.1 Extracting Inputs and Outputs

For each patient \( n \) in our \( N \) patient cohort, we extracted static observations and physiological data including labs and vital signs sampled hourly. All clinical variables were separately normalized to have zero mean and unit variance. Figure 1 shows how features are extracted for patients that are positive versus negative for in-hospital mortality.
Figure 1: Example positive and negative time-series to illustrate feature extraction. The two trajectories have input data extracted from time 0 to time of prediction $T = 24$. Figure inspired by Sherman et al [23].

**Static observations $S$.** Matrix $S : (N \times 8)$ contains 8 demographic variables for each patient $n$: their age at admission, gender, and other information about their ICU stay (their first ICU service type, and whether their admission was urgent, emergency, or elective).

**Per-timestep clinical observations $X$.** The clinical variable tensor $X : (N \times D \times T)$ contains $D = 28$ measurements of clinical variables for each of the $N$ patients at time $t$, discretized by hour. These 28 measurements consist of vital signs and labs: diastolic blood pressure, fio2, GCS score, heart rate, mean arterial blood pressure, systolic blood pressure, SRR, oxygen saturation, body temperature, urine output, blood urea nitrogen, magnesium, platelets, sodium, ALT and AST, hematocrit, po2, white blood cell count, bicarbonate, creatinine, lactate, pco2, glucose, INR, hemaglobin, and bilirubin. Missing values at timestep $t$ were imputed using either the most recent measurement of the variable, or the population median if the variable had not yet been measured during the patient’s stay. We use the patient’s first $T = 24$ hours of data to predict in-hospital mortality. We use subscripts to index into the tensor: for example, $X_t$ indicates the $(N \times D)$ matrix of measurements taken at time $t$.

**Per-timestep measurement indicators $M$.** The measurement indicator tensor $M : (N \times D \times T)$ contains indicator elements $M_{n,d,t}$ which are 1 if their corresponding clinical variable in $X_{n,d,t}$ was measured at time $t$, 0 otherwise.

**Outcome labels $y$.** Label vector $y$ contains indicators $y_n$ which are 1 if patient $n$ expired in-hospital, 0 otherwise.

### 4 Methods

Given a cohort of $N$ training examples $\{X, M, S, y\}$, we propose a novel procedure for learning predictive human-interpretable timeseries summaries. Concretely, we first compute summaries $H$ from clinical timeseries data $(X, M)$. We then use the summaries $H$ in addition to static data $S$ and clinical variables $X$ as input to a Logistic Regression model to predict labels $y$. This process of using human-interpretable summaries for prediction is shown in Figure 2.

In Section 4.1 we motivate and introduce our novel summary features. In Section 4.2 we give continuous relaxations of summary feature definitions to enable efficient inference of summary parameters. In Section 4.3 we discuss how predictive summary features can be jointly learned with downstream classification model parameters.

#### 4.1 Model: Defining human-interpretable summaries

Measures of central tendency and dispersion have commonly been used to summarize timeseries (see survey in [21]). Our key modelling innovations include adding additional features that correspond to how clinicians themselves describe timeseries. Our novel contributions include explicitly modelling the overall trend of a lab/vital and the number of hours that a lab/vital dips above or below a threshold, as well as allowing different features to be computed over different periods of time (e.g. the most recent 6 hours vs. the most recent 24 hours).
Figure 2: Illustrates summary extraction for prediction from timeseries data. First, non-static clinical variables \( \{X, M\} \) are used to compute interpretable summary features \( H \). Then, summary features \( H \), static features \( S \), and the non-static variables \( \{X, M\} \) at the time of prediction are given as input to a Logistic Regression model \( g \), which predicts output labels \( \hat{y} \). Figure inspired by Ghassemi, Szolovits et al \[24\].

The \( I = 13 \) summary statistics used in this study are listed in Table 2. Each of the summary statistics takes into account measurement indicators \( M \) so that clinical variable summaries are computed only using timesteps where the variable is measured. Each summary statistic is applied to each of the \( D \) clinical variables to create the summary feature tensor \( H : (N \times D \times I) \).

Below, we expand on the parameterization of our summary features. Next, we describe how we enable efficient, automated search over summary feature parameters. Importantly, our approach automates many processes associated with summary design, enabling optimization over summary parameters.

**Incorporating Duration.** Many prior works in clinical timeseries modelling do not use all timesteps for the patient, but instead only the most recent available data, such as the six or twelve hours before the time of prediction. In contrast to prior works, we explicitly model how much of each timeseries should be used to compute each of the \( I \) summaries in \( H \). For example, we may wish to exclude earlier measurements of a particular clinical variable if only the variable’s recent measurements before time of prediction \( T \) are significant for a prediction task. Specifically, for each variable \( d \) and for each summary function \( i \), we define a duration time \( C_{d,i} \). Only the variable’s timeseries data that occurred in the immediately previous \( C_{d,i} \) hours before the time of prediction is used to calculate summary \( i \). We organize all the duration time parameters \( C_{d,i} \) into a \((D \times I)\) matrix \( C \).

To exclude data that occurs before time \((T - C_{d,i})\) when computing summary features, we multiply each of the original timeseries variables by indicator variables for whether the measurements occurred within \( C_{d,i} \) hours before time of prediction \( T \). For example, a mean summary statistic would be computed using indicator variables \( 1(\cdot) \) as:

\[
H_{i=\text{mean}} : (N \times D) = \left( \sum_{t=1}^{T} 1(t > T - C_{i=\text{mean}}) \odot X_t \odot M_t \right) / \left( \sum_{t=1}^{T} M_t \odot 1(t > T - C_{i=\text{mean}}) \right)
\]

(1)

where \( \odot \) is the element-wise multiplication operator and division is performed element-wise. Our objective is to learn duration parameters \( C \) that maximize the predictiveness of their corresponding summary features \( H \).

**Threshold Parameters.** Some of the summary functions \( f_i \) in Table 2 have additional parameters such as thresholds. For example, one of the summaries is the proportion of the patient’s measured timeseries where their measured clinical variables are above some \( D \)-dimensional critical threshold parameter vector \( \phi^+ \) for each variable:

\[
\left( \sum_{t=1}^{T} M_t \odot 1(X_t > \phi^+) \right) / \left( \sum_{t=1}^{T} M_t \right)
\]

(2)

These summaries correspond to clinically-intuitive ideas, such as whether the patient been mostly well or sick. As with durations, we learn threshold parameters automatically to avoid burdening experts and to assist in prediction.
4.2 Continuous Relaxations for Efficient Inference

Learning Duration Parameters. In Equation\textsuperscript{1} we showed how summary functions \( f_i \) that only use the most recent \( C \) hours of data can be calculated using indicator variables \( 1(t > T - C_{d,i}) \). These indicator variables, however, do not have informative gradients and are not differentiable. To enable differentiable optimization for duration time parameters \( C \), we introduce weight parameters \( W \) by relaxing the indicator random variables using the sigmoid function \( \sigma(x) = \frac{1}{1 + e^{-x}} \). Using the duration parameter matrix \( C \), we define \( D \)-dimensional vectors \( w_{t,i} \) that compose weight tensor \( W : (T \times I \times D) \):

\[
w_{t,i} = \sigma((t - T + C_i)/\tau)
\]

For each feature \( d \) and summary \( i \), clinical observations \( X_{d,i} \) where \( t > T - C_{d,i} \) will have corresponding weights \( w_{t,i,d} \) near 1. Timesteps \( t \) where \( t < T - C_{d,i} \) will have corresponding weights \( w_{t,i,d} \) near 0. Temperature parameter \( \tau \) controls the harshness of the weight matrix: small temperatures push the sigmoid function towards its edges, learning weights that are closer to exactly 0 for timesteps before \( T - C_{d,i} \) and 1 for timesteps after \( T - C_{d,i} \), effectively functioning as the indicator variables in Equation\textsuperscript{1}.

Weighted summary functions \( f_i \) used to derive human-interpretable summaries \( \mathbf{H} \) can be found in Table\textsuperscript{2}. Duration parameters \( C \) that determine weight tensor \( W \) are included in \( \beta_H \), the set of all parameters necessary to compute the summaries.

<table>
<thead>
<tr>
<th>Description</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean of the time-series</td>
<td>( \left( \sum_{t=1}^{T} (w_{t,i} \odot X_t \odot M_t) \right) / \left( \sum_{t=1}^{T} M_t \odot w_{t,i} \right) )</td>
</tr>
<tr>
<td>Variance of the time-series</td>
<td>( \sigma \left( \frac{\sum_{t=1}^{T} M_t \odot w_{t,i}^{2} - \sum_{t=1}^{T} M_t \odot w_{t,i} \odot (X_t - \bar{X})^2}{\sum_{t=1}^{T} M_t \odot w_{t,i}^{2} - \sum_{t=1}^{T} M_t \odot w_{t,i}^{2}} \right) )</td>
</tr>
<tr>
<td>Indicator if feature was ever measured</td>
<td>( \sigma \left( \frac{\sum_{t=1}^{T} M_t \odot w_{t,i}^{2} - \sum_{t=1}^{T} M_t \odot w_{t,i} \odot (X_t - \bar{X})^2}{\sum_{t=1}^{T} M_t \odot w_{t,i}^{2} - \sum_{t=1}^{T} M_t \odot w_{t,i}^{2}} \right) )</td>
</tr>
<tr>
<td>Mean of the indicator sequence</td>
<td>( \left( \sum_{t=1}^{T} w_{t,i} \odot M_t \right) / \left( \sum_{t=1}^{T} w_{t,i} \right) )</td>
</tr>
<tr>
<td>Variance of the indicator sequence</td>
<td>( \left( \frac{\sum_{t=1}^{T} w_{t,i}^{2} - \sum_{t=1}^{T} w_{t,i} \odot (M_t - \bar{M})^2}{\sum_{t=1}^{T} w_{t,i}^{2}} \right) )</td>
</tr>
<tr>
<td># switches from missing to measured</td>
<td>( \min t \text{ s.t. } M_t = 1 )</td>
</tr>
<tr>
<td>First time the feature is measured</td>
<td>( \max t \text{ s.t. } M_t = 1 )</td>
</tr>
<tr>
<td>Last time the feature is measured</td>
<td>( \left( \sum_{t=1}^{T} w_{t,i} \odot M_t \odot \sigma \left( \frac{X_t - \bar{X}}{\tau} \right) \right) / \left( \sum_{t=1}^{T} M_t \odot w_{t,i} \right) )</td>
</tr>
<tr>
<td>Proportion of time above threshold ( \phi^+ )</td>
<td>( \left( \sum_{t=1}^{T} w_{t,i} \odot M_t \odot \sigma \left( \frac{X_t - \bar{X}}{\tau} \right) \right) / \left( \sum_{t=1}^{T} M_t \odot w_{t,i} \right) )</td>
</tr>
<tr>
<td>Proportion of time below threshold ( \phi^- )</td>
<td>( \left( \sum_{t=1}^{T} w_{t,i} \odot M_t \odot \sigma \left( \frac{X_t - \bar{X}}{\tau} \right) \right) / \left( \sum_{t=1}^{T} M_t \odot w_{t,i} \right) )</td>
</tr>
<tr>
<td>Slope of a L2 line</td>
<td>( \frac{\sum_{t=1}^{T} \sum_{w_{t,i} &lt; \bar{X} \text{ or } M_t &lt; 0} w_{t,i}(t - \bar{t}<em>w)(X_t - X_w)}{\sum</em>{t=1}^{T} \sum_{w_{t,i} &lt; \bar{X} \text{ or } M_t &lt; 0} w_{t,i}^2} )</td>
</tr>
<tr>
<td>Standard error of the L2 line slope</td>
<td>( 1 / \left( \sum_{t=1}^{T} w_{t,i} \odot (t - \bar{t}_w)^2 \right) )</td>
</tr>
</tbody>
</table>

Table 2: Table of functions \( f_i \) used to calculate human-interpretable summaries \( \mathbf{H} \). For each of the \( D \) clinical variables, all \( I \) of the above functions are applied to each of the \( N \) patients. Each of the \( I \) summary features \( i \) is defined with respect to \( D \)-dimensional weight vectors \( w_{t,i} \) defined in Section\textsuperscript{2}. Parameter \( \tau \) is a temperature parameter for the sigmoid function. \( 1(\cdot) \) denotes indicator variables for events inside the parentheses, \( \odot \) indicates element-wise matrix multiplication and division is done element-wise. Additionally, \( X = \sum_{t}^{T} M_t \odot X_t / \sum_{t}^{T} M_t \), \( \bar{M} = \frac{1}{T} \sum_{t}^{T} M_t \).

Learning Threshold Parameters. Our work relaxes summary definitions to enable differentiable optimization to learn summary parameters. The indicator variables used to define the proportion of hours that a patient’s timeseries is above thresholds \( \phi^+ \) in Equation\textsuperscript{2} are non-differentiable with respect to \( \phi^+ \). To enable differentiable optimization,
our work defines our threshold summary features using the sigmoid function $\sigma$ with temperature parameter $\tau$:

$$ f_{\text{threshold}}(X, M, W) = \left( \sum_{t=1}^{T} w_{t,i=\text{threshold}} \odot M_t \odot \sigma \left( \frac{X_t - \phi^+}{\tau} \right) \right) / \left( \sum_{t=1}^{T} M_t \cdot w_{t,i=\text{threshold}} \right) $$

(4)

Threshold parameters $\{\phi^+, \phi^-\}$ are included in $\beta_H$, the set of all parameters necessary to compute the summaries.

### 4.3 Learning Process

Our study uses summary features $H$, along with static variables $S$ and timeseries variables $\{X, M\}$ at prediction time $T = 24$ as input to a Logistic Regression model $g$ with coefficients $\beta_g$. Logistic Regression model $g$ outputs predicted probabilities $\hat{y} = p(y = 1|X, S, M)$. Our objective is to learn optimal summary and model parameters $\beta = \{\beta_H, \beta_g\}$. We jointly learn parameters $\beta$ by minimizing the loss function

$$ L(\beta; X, M, S, y) = -\frac{1}{N} \sum_{n=1}^{N} \omega_n \left( y_n \cdot \log[g(X_n, M_n, S_n, \beta)] + (1 - y_n) \cdot \log[1 - g(X_n, M_n, S_n, \beta)] \right) + \Omega(\beta_g) $$

(5)

Our loss function is the sum of the weighted binary cross-entropy loss using predictive model $g$ and the regularization penalty $\Omega(\beta_g)$. To account for class imbalance, we reweight each training example $n$’s loss contribution by $\omega_n$, the inverse of its class frequency in the training dataset.

**Horseshoe Regularization on Coefficients.** We explicitly optimize for sparsity in model parameters $\beta_g$ via our regularization penalty in our training objective. We use a Horseshoe regularization penalty $\Omega(\beta_g)$ with shrinkage parameter $1$ to encourage sparsity in the learned regression coefficients $[25]$.

### 5 Experiments

We compare models learned using our summaries to other interpretable models as well as deep learning baselines. We show that our models outperform other traditional human-interpretable model classes and achieve performance comparable to deep models on the in-hospital mortality task.

**Model configurations.** Our baseline models are: Ridge Logistic Regression models that take as input only the patient timeseries measured at the time of prediction $T$, and Ridge Logistic Regression and LSTM models that take as input all of the patient timeseries. We used an LSTM as our deep baseline architecture as prior works document their superior performance at mortality prediction from clinical timeseries data $[26]$.

Our LSTM models were trained on the sequential timeseries data $\{X, M\}$ with a step size of 1 hour. The LSTM hidden states at each timestep were then used to predict both the next timestep of the patient timeseries $X$, and to predict outcome labels $y$. All $T$ of the output hidden states $X_t$ were input to a fully-connected layer, which output predictions of the next timestep $X_{t+1}$. The last output hidden state at time $T$ was also input with static features $S$ to a fully-connected layer to predict outcome labels $y$. The LSTM models were trained to minimize both the mean-squared error of the next state prediction, and the binary cross entropy loss of the classification prediction. ReLU activation functions were applied to both fully-connected layers.

**Training Details.** For training and testing, we split the cohort of $N$ patients into train and test sets, where all data associated with each patient is either in train or test. All performance metrics are averaged across five train-test splits. Ridge baseline models were implemented using RidgeCV from scikit-learn $[27]$, LSTMs as well as our summary-based Logistic Regression models were implemented with PyTorch, and trained with the Adam optimizer $[28]$ at a batch size of $256$. We trained all of our Logistic Regression models for $30,000$ epochs and LSTM models for $10,000$ epochs using early stopping. All hyperparameters (including the LSTM hidden state and layer dimensions, optimizer learning rate, and regularization parameters) were selected via random hyperparameter search $[29]$. All temperature parameters $\tau$ were set to $0.1$.

Our final learned LSTM models have hidden state dimension $32$, $1024$ nodes in the layer to predict the next timestep, and $64$ nodes in the layer to predict labels $y$. They are trained using a learning rate of $1e^{-05}$. Our final learned models with summaries use $\alpha = 1e^{-05}$, Horseshoe shrinkage parameter $1.0$, and learning rate $1e^{-05}$. 

586
6 Results

Our learned models achieve performance comparable to state-of-the-art baselines. Table 3 compares the performance of our learned models with linear and deep baselines for the in-hospital mortality prediction task. Applying a linear model to the learned summary features $H$ consistently improves AUCs in comparison to using a linear model on clinical timeseries and static data alone. Our models achieve an AUC performance comparable to state-of-the-art LSTM models with test AUCs of $0.9000 \pm 0.0223$. Notably, our models that allow differentiable optimization to learn duration times outperform models that compute all summary features using the entire duration of the patient timeseries. This implies that there is predictive value in explicitly modelling how much of each variable’s clinical timeseries should be considered for a specific prediction task.

<table>
<thead>
<tr>
<th>Model</th>
<th>Train set AUC</th>
<th>Test set AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR trained on ${S, M}$ and $X$ at time $T$ only</td>
<td>0.8653 ± 0.0013</td>
<td>0.8626 ± 0.0079</td>
</tr>
<tr>
<td>LR trained on ${S, M, X}$</td>
<td>0.8931 ± 0.0015</td>
<td>0.8668 ± 0.0122</td>
</tr>
<tr>
<td>LSTM trained on ${S, M, X}$</td>
<td>0.9101 ± 0.0018</td>
<td>0.9000 ± 0.0223</td>
</tr>
<tr>
<td>Our model, trained on ${S, M, X}$, non-differentiable durations $C$</td>
<td>0.9065 ± 0.0019</td>
<td>0.8818 ± 0.0063</td>
</tr>
<tr>
<td>Our model, trained on ${S, M, X}$, differentiable durations $C$</td>
<td>0.9074 ± 0.0016</td>
<td>0.8867 ± 0.0061</td>
</tr>
</tbody>
</table>

Table 3: Performance of learned models on the in-hospital mortality prediction task. AUCs are averaged over five train-test splits with their standard error. Abbreviations: LR, Logistic Regression; LSTM, long short-term memory. $C$ refers to the duration parameters defined in Section 4.2.

Our learned models use fewer features to achieve higher accuracy in comparison to other interpretable baselines. To evaluate the sparsity of each model, we performed a set of ablation experiments where we zeroed all but the $N$ coefficients with the largest magnitudes for each of the learned Logistic Regression models. In Figure 3, we show the average test set AUC for Logistic Regression baseline versus our models using only $N$ coefficients. Our models consistently outperform baseline models when all but $N$ coefficients are zeroed, suggesting that our models learn a smaller and more predictive set of important features.

Our learned models are interpretable. Table 4 shows 15 key summary features that consistently have the largest learned Logistic Regression coefficients across train-test splits. The corresponding coefficients for each feature can be interpreted as a measure of the feature’s contribution to the final classification label. For example, because the mean of the patient’s GCS has a large negative coefficient, this means that patients with higher mean GCS scores will be assigned lower predicted probabilities for in-hospital mortality. Therefore our models are decomposable [30], as each
of the model’s features and coefficients has an intuitive clinical explanation.

**Our learned summary features are clinically sensible.** The vast majority of the key summary features learned by our models shown in Table 4 are supported by studies in medical literature. For instance, it is widely accepted that patients who are older tend to have lower chances of survival in ICU settings [31, 32]. Similarly, patients with lower GCS scores of below 6 tend to have severe injuries and higher chances of mortality [33]. Notably a lower GCS score in the later hours of a patient’s hospitalisation significantly reduces a patient’s chances of survival [34, 35]. Finally, the normal range of features such as the blood oxygen saturation (SPO$_2$) is between 95% and 100%. An SPO$_2$ consistently below 90% indicates hypoxaemia or potential respiratory distress. These patients have to be mechanically ventilated in ICU and frequently have lower chances of survival [36, 37].

<table>
<thead>
<tr>
<th>Feature</th>
<th>Aggregation</th>
<th>Time</th>
<th>Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>static value</td>
<td>-</td>
<td>14.9</td>
</tr>
<tr>
<td>BUN</td>
<td>value at hour 24</td>
<td>6.2</td>
<td></td>
</tr>
<tr>
<td>GCS</td>
<td>mean over hours 5 - 24</td>
<td>- 5.59</td>
<td></td>
</tr>
<tr>
<td>HR</td>
<td>value at hour 24</td>
<td>4.69</td>
<td></td>
</tr>
<tr>
<td>FiO$_2$</td>
<td>value at hour 24</td>
<td>4.31</td>
<td></td>
</tr>
<tr>
<td>Hct</td>
<td>times measured over hours 2 - 24</td>
<td>- 3.81</td>
<td></td>
</tr>
<tr>
<td>HR</td>
<td>mean over hours 2 - 24</td>
<td>3.78</td>
<td></td>
</tr>
<tr>
<td>GCS</td>
<td>value at hour 24</td>
<td>- 3.62</td>
<td></td>
</tr>
<tr>
<td>GCS</td>
<td>hours below 6.08</td>
<td>hours 7 - 24</td>
<td>3.37</td>
</tr>
<tr>
<td>Creatinine</td>
<td>hours below 0.35 mg/dL</td>
<td>hours 5 - 24</td>
<td>2.76</td>
</tr>
<tr>
<td>FiO$_2$</td>
<td>hours above 62.96%</td>
<td>hours 16 - 24</td>
<td>2.59</td>
</tr>
<tr>
<td>SpontaneousRR</td>
<td>mean over hours 2 - 24</td>
<td>2.29</td>
<td></td>
</tr>
<tr>
<td>GCS</td>
<td>times measured over hours 5 - 24</td>
<td>2.23</td>
<td></td>
</tr>
<tr>
<td>Sodium</td>
<td>hours below 131.57 mEq/L</td>
<td>hours 1 - 24</td>
<td>2.16</td>
</tr>
<tr>
<td>WBC</td>
<td>hours below 0.78 cells/mL</td>
<td>hours 6 - 24</td>
<td>2.10</td>
</tr>
<tr>
<td>SPO$_2$</td>
<td>hours below 92.36%</td>
<td>hours 10 - 24</td>
<td>2.04</td>
</tr>
</tbody>
</table>

Table 4: Key summary features, sorted from largest to smallest coefficient magnitudes, from learned models.

**Initialization sensitivity.** In general, we observed that our optimization procedure is stable, learning the same 15 key summary features across different stochastic parameter initializations and train-test splits. However, there are cases where we observed that the learned duration parameters $C$ varied depending on their initialization. As such, we recommend that practitioners incorporate prior knowledge about the clinical prediction task when initializing the duration time parameters. For example, if examining the entire duration of the patient’s timeseries is necessary for a prediction task, then the duration parameters should be initialized to include the entire timeseries by default.

7 Discussion & Conclusion

In this work, we defined functions to compute interpretable, parameterizable summaries of clinical timeseries, and developed relaxations so that our summary parameters could be jointly learned with a downstream predictive model. In our experiments, we used Logistic Regression to make predictions because its coefficients are easily decomposable [30]. However, because our learned summaries are inherently interpretable, any other interpretable architecture could be used instead. Our methodology is generalizable and enables the efficient learning of intuitive and predictive timeseries summaries without placing any assumptions on the downstream model architecture.

**Future work.** Our study poses many interesting directions for future work. One avenue would be to conduct a user study to validate the human-interpretability and decomposability of our proposed summary features. Another would be to evaluate whether the summary features learned for particular critical care prediction tasks remain predictive for a wider set of critical care prediction tasks. Finally, we could also develop additional summary statistic functions, or expand our framework to consider sharing duration parameters across features or across summaries to better model dependencies between clinical labs and vitals—as many physiological events are characterized by several simultaneous
changes to multiple labs and vitals.  

**Conclusion.** In this paper, we propose a new method to learn interpretable and predictive summary features from clinical timeseries data. In addition to introducing novel summary statistics including slope and threshold features, our work differs from prior work by learning the duration of timeseries data that should be used to compute each summary. We demonstrate that our learned timeseries summaries achieve performance quality comparable to state-of-the-art deep models when trained to predict early patient mortality risk on real patient data. We also qualitatively validate our models to confirm their interpretability and sensibility. Our work is an important step towards optimizing for representations of clinical timeseries data that are both highly predictive and interpretable.

**Acknowledgements:** NJ and FDV acknowledge support from NIH R01 MH123804-01A1. SP acknowledges support from the Miami Foundation and SNSF P2BSP2-184359.

**References**


Unsupervised characterization of Major Depressive Disorder medication treatment pathways

Barrett Jones, MA¹, Colin G. Walsh, MD, MA¹,²,³
¹Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, Tennessee, USA; ²Department of Medicine, Vanderbilt University Medical Center, Nashville, Tennessee, USA; ³Department of Psychiatry, Vanderbilt University Medical Center, Nashville, Tennessee, USA

Abstract

Learning health systems have the ability to systematically evaluate treatments and treatment pathways. Characterization of treatment pathways can enhance a health system’s ability to perform systematic evaluation to improve care quality. In this study we use a Long-Short Term Memory (LSTM) autoencoder model to systematically characterize treatment pathways in a prevalent phenotype—Major Depressive Disorder (MDD). LSTM autoencoder models generate representations of medication treatment pathways that account for temporality and complex interactions. Patients with similar pathways are grouped with K-means clustering. Clusters are characterized by analysis of medication utilization sequences and trends, as well as clinical features, such as demographics, outcomes and comorbidities. Cluster characterization identifies endotypes of MDD including acute MDD, moderate-chronic MDD and severe-chronic, but managed MDD.

Introduction

Learning health systems are those that work in a continuous cycle converting data to knowledge to process. These systems have the capability to systematically evaluate treatments from observational data and augment processes to improve system performance¹–³. An important step towards systematic evaluation is characterization of in place treatment pathways⁴. In this study, we perform a systematic characterization of treatment pathways of Major Depressive Disorder (MDD) from observational electronic health record (EHR) data at Vanderbilt University Medical Center (VUMC). Clinical trials have shown the effectiveness of antidepressants in treatment of MDD relative to placebo⁴, but there is a need for better understanding and representation of long term treatment practices⁵–⁷. Efforts to standardize treatment pathways have resulted in decreased variance in treatment of MDD over time, but heterogeneity of treatment pathways remains between institutions⁵. Due to the prevalence of MDD and impact on daily life, characterization of treatment pathways in MDD may be beneficial to healthcare organizations in improving outcomes and quality of care⁶–⁹.

Past studies have modeled treatment paths by visualizing medication sequences²,⁸. We built upon these studies by modeling medication treatment pathways in the VUMC Synthetic Derivative¹⁰ accounting for temporality as well as treatment resistant depression (TRD). TRD is defined as failure to respond to two or more antidepressant trials¹¹. TRD is of particular interest in this study for the potential use of treatment pathways in prediction of TRD in future studies. In this study, electroconvulsive therapy (ECT) was used as a surrogate for TRD¹².

Long-Short Term Memory (LSTM) neural networks are deep learning algorithms that have the ability to model temporal data with complex relationships like those characteristic of medication data. An LSTM is a form of Recurrent Neural Network that stores information over extended time intervals and employs a gating method to address the exploding gradient problem found in some Recurrent Neural Network applications¹³. Autoencoders have the ability to generate simplified encodings of complex data structures. Autoencoders are composed of two sub-models: an encoder and decoder. The encoder receives an input data set and reduces the dimensionality to a hidden layer. The decoder takes as input the hidden layer from the encoder model and reconstructs the input data¹⁴,¹⁵. The hidden layer generated by the encoder model is a denoised continuous vector representation of the model input. LSTM autoencoders have been used to learn representations for video sequences¹⁶ and for biomedical endotyping¹⁷. In this study, we hypothesize that LSTM autoencoders can effectively represent medication treatment pathways and that characterization of encoding clusters will allow for differentiation of clusters by medication treatment pathways and clinically relevant variables.
Methods

Data Description

The patient cohort for this study included patients aged 18 to 90, diagnosed with major depressive disorder (MDD) seen at least two times, six months apart, at Vanderbilt University Medical Center (VUMC) located in the United States Mid-South in Nashville, TN. Patients with MDD were identified that had at least one of the following ICD-9 codes: 311.x, 296.2x, 296.3x, 300.4x or ICD-10 codes: F32.xx, F33.xx, F34.1 and their index diagnosis took place prior to 12/31/2016 to ensure a minimum of three years of medication data. We excluded any patients with a diagnosis of Bipolar Disorder or Schizophrenia and those that were not prescribed an antidepressant (see Table 5) within three years of diagnosis.

Medications were extracted for three years following first MDD diagnosis and subset to include only prescriptions. Because the population with TRD was of particular interest for downstream prediction of TRD status before it occurred, the cohort with MDD who received ECT, a surrogate for TRD in some studies, had medication data censored one day prior to ECT treatment date. Medications were grouped using the World Health Organization Anatomical Therapeutic Chemical (ATC) classification level 5. Medications included in the study are listed in the Appendix Table 5.

Unsupervised Feature Learning

Medication data were converted to a categorical quarterly time series with indicators for each medication class if observed during a quarter. The resulting dataset had dimension N x M x T, with N the number of patients, M the number of medication classes, and T the number of quarters in our observation window.

An LSTM autoencoder was fit on the medication data to generate a continuous-multidimensional representation of each patient’s medication pathway. The model was built using the Keras Python library. We tested multiple model structures in order to tune the number of encoder layers as well as the dimension of the encoder layers. We performed a grid search on 15 candidate models with one to three encoder layers and final encoder layer dimensions of 8, 16, 24, 32 and 64. Models were evaluated based on reconstruction mean squared error (MSE) on a 25% validation test set. Model selection was based on visual inspection of plots with encoding dimension on the x axis and MSE on the y. The model with an MSE such that increasing encoding dimension does not meaningfully reduce MSE was selected for clustering.

The encodings of the selected model were clustered using the K-means algorithm. The Elbow Method was utilized for selection of K—the number of clusters. In this method the sum squared distance from the cluster centroid is plotted against K and the “elbow” is selected by visual inspection as the point at which increasing K does not meaningfully reduce sum squared distance from the centroid.

Cluster Characterization

We characterized clusters by performing an analysis of demographics, clinical outcomes, comorbidities, and utilization trends. Demographics include gender, age and race. The clinical outcomes measured in the study are all cause mortality, admission, ER Visit and a TRD surrogate (electroconvulsive therapy [ECT]) within the three-year period of analysis. Comorbidities are identified using Agency for Healthcare Research and Quality (AHRQ) Clinical Classification Software (CCS) to map ICD 9 and 10 codes to clinically meaningful comorbidities. We perform a Chi-square test of association between clusters and categorical demographics—gender and race—and perform ANOVA for continuous variables—age and comorbidity count—with post-hoc Bonferroni corrected confidence intervals to measure differences between clusters. Mental health comorbidity association with clusters were also analyzed by Chi-square with post-hoc Bonferroni corrected significance tests to identify clusters with enriched comorbidity prevalence.

Medication and visit utilization analysis included: calculating the mean number of prescriptions per patient, mean unique number of prescription classes per patient and single medication treatment rates—the proportion of patients prescribed a single medication class for the entire period of analysis. Trend graphs visualized three-quarter rolling averages of medication utilization. By cluster medication sequences were visualized by sunburst plots. For
visualization purposes medications were combined into clinically meaningful groupings as defined by physicians from the VUMC department of psychiatry. These groupings are available in Table 5 included in the appendix of this paper. Additionally, we account for instances in which patients are prescribed multiple medications in a short time period. We will refer to these cases as Multi-medications-therapy and define it as any instance in which a patient is prescribed two different medications (ATC 5 level) within a five-day period.

**Results**

**Model Selection**

Autoencoder models were fit and validation MSE was plotted by encoding dimension for each of the model structures in Figure 1. The three-layer encoder model outperforms the one and two encoder layer models at each candidate encoding dimension. Validation MSE in the three-layer encoder model is 0.0034 with encoding dimension equal to 24. At dimension 32 and 64 the validation MSE is 0.0032 and 0.0020 respectively. As can be seen in Figure 1, the rate at which the three-layer encoder model MSE (green) is decreasing is reduced at dimension 24. The model with three encoding layers and encoding layer dimension of 24 is selected for our final model. Encodings from the selected model were cluster using the K-means algorithm. The elbow method was utilized for selection of K—the number of clusters. Visual inspection of the Figure 1 (right) indicates that the “elbow” occurs at K=5.

![Figure 1. Autoencoder and K-means clustering model selection plots.](image)

The reconstructed MSE for each candidate model is plotted with lines colored by the number of encoding layers (left). The within cluster sum squares are plotted for the K-means algorithm with K=1 through K=14 (right).

**Cluster Characterization**

Summary statistics by cluster were calculated for demographics, outcomes and comorbidities. 46,454 patients met the inclusion criteria for this study. Five clusters were identified, cluster 2 being the largest (n=13,908, 29.1%). ANOVA and Chi-square tests were conducted for association between clusters and variables of interest. For each of the variables tested differences between clusters were found to be statistically significant. Post-hoc Bonferroni corrected confidence interval margin of errors are included in Table 1. The majority of patients in the study are female (66.4%) and Cluster 4 is higher proportion female (70.4%, 95% CI = 69.4%-71.4%). Cluster 5 has elevated mortality (3.6%, 95% CI = 3.3%-3.9%) and Cluster 2 has elevated admissions (66.6%, 95% CI = 66.0%-67.2%) and ECT (0.56%, 95% CI = 0.46%-0.66%). Cluster 3 has the highest CCS comorbidity count on average (17.6, 95% CI = 17.3-17.6).
Table 1. Clinical characteristics by cluster. For each metric in the table the margin of error for a Bonferroni corrected 95% confidence interval is included in parentheses.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>P-value</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demographics N</td>
<td>6,238</td>
<td>13,908</td>
<td>10,616</td>
<td>5,112</td>
<td>10,580</td>
<td>&lt;0.001</td>
<td>46,454</td>
</tr>
<tr>
<td>% Female</td>
<td>68.2 (0.9)%</td>
<td>64.5 (0.6)%</td>
<td>70.4 (0.7)%</td>
<td>65.1 (1.0)%</td>
<td>61.5 (0.7)%</td>
<td>&lt;0.001</td>
<td>66.4%</td>
</tr>
<tr>
<td>Mean Age</td>
<td>48.8 (0.6)</td>
<td>45.8 (0.4)</td>
<td>48.5 (0.5)</td>
<td>47.6 (0.6)</td>
<td>47.1 (0.5)</td>
<td>&lt;0.001</td>
<td>47.3</td>
</tr>
<tr>
<td>Race/Ethnicity Black/Non-Hispanic</td>
<td>9.6 (0.6)%</td>
<td>10.3 (0.5)%</td>
<td>8.6 (0.5)%</td>
<td>10.9 (0.8)%</td>
<td>8.7 (0.5)%</td>
<td>&lt;0.001</td>
<td>9.5%</td>
</tr>
<tr>
<td>Black/Hispanic-Latino</td>
<td>0.0 (0.0)%</td>
<td>0.1 (0.1)%</td>
<td>0.1 (0.1)%</td>
<td>0.1 (0.1)%</td>
<td>0.1 (0.1)%</td>
<td>&lt;0.001</td>
<td>0.1%</td>
</tr>
<tr>
<td>White/Non-Hispanic</td>
<td>85.6 (0.8)%</td>
<td>83.7 (0.6)%</td>
<td>84.9 (0.6)%</td>
<td>85.6 (0.6)%</td>
<td>84.9 (0.6)%</td>
<td>&lt;0.001</td>
<td>85.3%</td>
</tr>
<tr>
<td>White/Hispanic-Latino</td>
<td>1.2 (0.2)%</td>
<td>1.7 (0.3)%</td>
<td>1.2 (0.3)%</td>
<td>1.5 (0.3)%</td>
<td>1.5 (0.3)%</td>
<td>&lt;0.001</td>
<td>1.5%</td>
</tr>
<tr>
<td>Other</td>
<td>3.7 (0.2)%</td>
<td>4.3 (0.2)%</td>
<td>3.0 (0.2)%</td>
<td>2.9 (0.2)%</td>
<td>4.0 (0.2)%</td>
<td>&lt;0.001</td>
<td>3.7%</td>
</tr>
<tr>
<td>Outcomes Mortality %</td>
<td>2.4 (0.3)%</td>
<td>1.5 (0.2)%</td>
<td>2.2 (0.2)%</td>
<td>1.3 (0.3)%</td>
<td>3.6 (0.3)%</td>
<td>&lt;0.001</td>
<td>2.2%</td>
</tr>
<tr>
<td>Admission %</td>
<td>42.1 (1.0)%</td>
<td>66.6 (0.6)%</td>
<td>45.1 (0.8)%</td>
<td>42.6 (1.1)%</td>
<td>47.9 (0.8)%</td>
<td>&lt;0.001</td>
<td>51.5%</td>
</tr>
<tr>
<td>ER Visit %</td>
<td>28.9 (0.9)%</td>
<td>27.9 (0.6)%</td>
<td>28.5 (0.7)%</td>
<td>29.5 (1.0)%</td>
<td>27.3 (0.7)%</td>
<td>0.024</td>
<td>28.2%</td>
</tr>
<tr>
<td>ECT %</td>
<td>0.24 (0.1)%</td>
<td>0.56 (0.1)%</td>
<td>0.20 (0.07)%</td>
<td>0.17 (0.09)%</td>
<td>0.27 (0.08)%</td>
<td>&lt;0.001</td>
<td></td>
</tr>
<tr>
<td>Comorbidities Mean CCS</td>
<td>16.0 (0.3)</td>
<td>13.3 (0.3)</td>
<td>17.6 (0.3)</td>
<td>15.6 (0.3)</td>
<td>16.7 (0.3)</td>
<td>&lt;0.001</td>
<td>15.7</td>
</tr>
</tbody>
</table>

Medication utilization statistics by cluster are included in Table 3 and trend plots of medication and visit utilization in Figure 2. Clusters 2 and 4 have the lowest mean count of prescriptions (3.0, 3.6) during the period of analysis and the lowest number of unique prescriptions by medication class (1.5, 1.4) as well as the highest rates of single medication treatment (66.4%, 68.9%). Clusters 3 and 5 have the highest prescription counts (12.4, 9.8) and the lowest rates of single medication treatment (31.7%, 37.2%).

Table 2. Medication utilization summary statistics by cluster.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>6,238</td>
<td>13,908</td>
<td>10,616</td>
<td>5,112</td>
<td>10,580</td>
</tr>
<tr>
<td>Mean prescription count</td>
<td>5.5</td>
<td>3.0</td>
<td>12.4</td>
<td>3.6</td>
<td>9.8</td>
</tr>
<tr>
<td>Mean unique prescription class count (ATC Level 5)</td>
<td>1.7</td>
<td>1.5</td>
<td>2.3</td>
<td>1.4</td>
<td>2.2</td>
</tr>
<tr>
<td>Single Medication Treatment Rates</td>
<td>58.5%</td>
<td>66.4%</td>
<td>31.7%</td>
<td>68.9%</td>
<td>37.2%</td>
</tr>
</tbody>
</table>
Figure 2. Medication and visit utilization trends. Three quarter moving average plots of the percentage of patients with a prescription (left) and visit (right) during each quarter of the period of analysis by cluster.

The most common initial medication type prescribed was selective serotonin reuptake inhibitors (SSRI) in the patients studied (50.1%). This ranged from 47.7% in cluster 4 up to 51.4% in cluster 3. For patients receiving an SSRI initially, the majority of these patients did not receive another prescription within the period of analysis (66.2%) and 9.1% were prescribed a SSRI as their second antidepressant. In our study population 18.4% of patients received multi-medicating-therapy. Rates of multi-medicating-therapy range between 11.5% in cluster 1 and 22.6% in cluster 2.
Figure 3. Sunburst plot of medication sequences by cluster. The first (inner) level of the plot represents the distribution of initial antidepressant types with the second level representing the second antidepressant type prescribed and so on.

The Chi-square analysis of association between mental health comorbidities and clusters was conducted to identify comorbidity-cluster pairs that deviated from expected frequency given the overall comorbidity frequency within the population and the cumulative comorbidity burden within the cluster. Results of the analysis are included in Table 4.

Table 3. Chi-square analysis of mental health comorbidities by cluster. Chi-square analysis to test association between cluster and mental health comorbidities. Post-hoc Bonferroni corrected significance tests of the Chi-square statistic are performed for each comorbidity-cluster pair at the 90%, 95%, and 99% significance levels (indicated by color).
Discussion

Overall, we find that the unsupervised learning methods are able to separate the study population into clusters using medication data alone. We find differences in medication utilization, medication heterogeneity, temporal trends, as well as statistically significant differences between clusters in clinical features. This work builds on the characterization of treatment pathways research\cite{2,3,8} by leveraging machine learning methods that account for temporality and extending characterization to analyze relevant clinical features associated with treatment pathways.

Cluster characterization analysis examined three dimensions to describe the autoencoder and clustering outputs: utilization, medication heterogeneity, and temporal trends. Table 2 displays both utilization (mean prescription count) and medication heterogeneity (mean unique prescriptions, single medication treatment rates). The medication sequences in the sunburst plots (Figure 3) visualize medication heterogeneity between clusters and the quarterly moving average plots of percent patients with prescription and visit provide insight into temporal trends (Figure 2). In differentiating between clusters, medication utilization trends are a key factor. The profiles of clusters 3 and 5 are similar in aggregate medication utilization statistics, as well as medication sequences. However, medication utilization trends downward initially after diagnosis, but utilization sustains a consistent level in quarters 6 through 12 from MDD diagnosis. Conversely, cluster 5 medication utilization trends downward throughout the period of analysis.

In addition, we describe the clusters by clinical features such as demographics, comorbidities and outcomes. Differences between each of the demographics measured are found to be statistically significant between clusters. Outcomes measured include mortality, admission, ER visits, and ECT. Themes emerge from characterization analysis. For example, cluster 2 is characterized by higher rates of substance use disorders and self-harm as well as higher provision of ECT. These factors suggest this cluster captures high acuity psychiatric patients with depression at our medical center. This finding is also supported by medication treatment patterns, cluster 2 has high initial medication utilization and the highest rate of multi-medications treatment among the clusters. Clusters 1 and 4 are characterized by low utilization, outcomes and comorbidity profile. We suggest patients in these clusters have a moderate severity chronic MDD. Clusters 3 and 5 convey high utilization, and high comorbid burden with relatively low rates of substance abuse and self-harm. We assert clusters 3 and 5 are patients with severe, chronic, but managed MDD.

Our study is limited to analysis of data from a single institution. Our institution is an open system with finite specialty mental health access: patients included in our study might receive care at an institution other than VUMC or the patients in this study might be incident users of antidepressants\cite{21}. We seek to limit this risk by requiring two visits to VUMC at least six months apart. Additionally, there are limitations in representation of the medication data. Our methods seek to address some of those limitations by representing medications as a time-series and leveraging an LSTM auto-encoder. In this work, we do not explicitly seek to capture concurrent medications. However, concurrent medications may be implicitly captured to some degree by the quarterly time series representation. In this study we limit to including only medications from VUMC at the time they are prescribed. Information about duration of prescription, adherence, or dose changes are not included in our model but remain germane to effective treatment of MDD.

Future work should seek to expand treatment pathway characterization to additional health systems and to analyze additional phenotypes. In addition, health information exchanges can provide a more complete picture of the patient health record\cite{22}.

Conclusion

We clustered the output of an LSTM autoencoder on time-series of antidepressant prescriptions in patients diagnosed with MDD. We identified common, clinically relevant patterns in prescribing practices. Cluster characterization established that prescribing practices differ on multiple dimensions: utilization, medication heterogeneity, temporal trends. Clusters are found to be associated with a clinically meaningful features including demographics, outcomes, and comorbidities. This method provides insight into endotypes of MDD.
**Appendix**

**Table 4.** Medications included in the study by Anatomical Therapeutic Chemical classification (ATC) and clinical grouping

<table>
<thead>
<tr>
<th>Clinical Grouping</th>
<th>Medication Name (ATC Level 5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selective serotonin reuptake inhibitors (SSRI)</td>
<td>citalopram</td>
</tr>
<tr>
<td></td>
<td>fluoxetine</td>
</tr>
<tr>
<td></td>
<td>paroxetine</td>
</tr>
<tr>
<td></td>
<td>escitalopram</td>
</tr>
<tr>
<td></td>
<td>fluvoxamine</td>
</tr>
<tr>
<td></td>
<td>sertraline</td>
</tr>
<tr>
<td>Tricyclic antidepressants</td>
<td>doxepin</td>
</tr>
<tr>
<td></td>
<td>amitriptyline</td>
</tr>
<tr>
<td></td>
<td>desipramine</td>
</tr>
<tr>
<td></td>
<td>nortriptyline</td>
</tr>
<tr>
<td></td>
<td>protriptyline</td>
</tr>
<tr>
<td></td>
<td>clomipramine</td>
</tr>
<tr>
<td></td>
<td>maprotiline</td>
</tr>
<tr>
<td></td>
<td>amoxapine</td>
</tr>
<tr>
<td></td>
<td>imipramine</td>
</tr>
<tr>
<td>Monoamine oxidase inhibitors, non-selective (MAOI)</td>
<td>tranylcypromine</td>
</tr>
<tr>
<td></td>
<td>phenelzine</td>
</tr>
<tr>
<td></td>
<td>isocarboxazid</td>
</tr>
<tr>
<td>Second generation antipsychotic medications</td>
<td>aripiprazole</td>
</tr>
<tr>
<td></td>
<td>quetiapine</td>
</tr>
<tr>
<td></td>
<td>olanzapine</td>
</tr>
<tr>
<td></td>
<td>ziprasidone</td>
</tr>
<tr>
<td>Other antidepressants</td>
<td>vilazodone</td>
</tr>
<tr>
<td></td>
<td>bupropion</td>
</tr>
<tr>
<td></td>
<td>vortioxetine</td>
</tr>
<tr>
<td></td>
<td>trazodone</td>
</tr>
<tr>
<td></td>
<td>mirtazapine</td>
</tr>
<tr>
<td></td>
<td>nefazodone</td>
</tr>
<tr>
<td>Centrally acting sympathomimetics</td>
<td>methylphenidate</td>
</tr>
<tr>
<td></td>
<td>dexamfetamine</td>
</tr>
<tr>
<td>Mood Stabilizers</td>
<td>lithium</td>
</tr>
<tr>
<td></td>
<td>lamotrigine</td>
</tr>
<tr>
<td>Serotonin-norepinephrine reuptake inhibitors (SNRI)</td>
<td>desvenlafaxine</td>
</tr>
<tr>
<td></td>
<td>duloxetine</td>
</tr>
<tr>
<td></td>
<td>venlafaxine</td>
</tr>
<tr>
<td></td>
<td>milnacipran</td>
</tr>
</tbody>
</table>

**References**


**Acknowledgements**

We thank Drs. Warren Taylor and Patricia Andrews (Vanderbilt University Department of Psychiatry) for providing guidance on the medications included in the study, defining the clinical groupings of medications, and providing guidance on clinical interpretation of results.
Gene-Based Analysis Reveals Sex-Specific Genetic Risk Factors of COPD

Jaehyun Joo, Ph.D.1, Blanca Himes, Ph.D.1
1Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

Abstract

Sex-specific differences have been noted among people with chronic obstructive pulmonary disease (COPD), but whether these differences are attributable to genetic variation is poorly understood. The availability of large biobanks with deeply phenotyped subjects such as the UK Biobank enables the investigation of sex-specific genetic associations that may provide new insights into COPD risk factors. We performed sex-stratified genome-wide association studies (GWAS) of COPD (male: 12,958 cases and 95,631 controls; female: 11,311 cases and 123,714 controls) and found that while most associations were shared between sexes, several regions had sex-specific contributions, including respiratory viral infection-related loci in/near C5orf56 and PEL11. Using the newly developed R package ‘snpsettest’, we performed gene-based association tests and identified gene-level sex-specific associations, including C5orf56 on 5q31.1, CFDP1/TMEM170A/CHST6 on 16q23.1 and ASTN2/TRIM32 on 9q33.1. Our results identified promising genes to pursue in functional studies to better understand sexual dimorphism in COPD.

Introduction

Chronic obstructive pulmonary disease (COPD) is a public health challenge worldwide1. Smoking is the most important COPD risk factor, accounting for 8 out of 10 COPD-related deaths in the U.S., but the relationship is complex and many smokers do not develop COPD1,2. In the past, COPD was perceived as a disease affecting mostly older men, but recent evidence shows that its prevalence has increased faster in women than men due in part to the increased rate of smoking among women3,4. Women may be more susceptible to the harmful effects of cigarette smoke, as some studies have found that women have worse disease outcomes for the equivalent quantity of cigarettes consumed compared to men, although research on this topic is not conclusive5–7. An observational study found that as women who smoked aged, they had an accelerated decline in percent predicted forced expiratory volume in one second (FEV1) and a faster recovery in lung function after smoking cessation than men8. In smoking- or FEV1-matched men and women, women with COPD reported more symptoms, including dyspnea and cough5,6,7,9. Among people with severe COPD, women had less extensive emphysema and thicker small airway walls relative to luminal perimeters than men6. The mechanisms underlying these sex-specific differences in COPD are largely unknown, but they may be influenced by genetic factors. Genome-wide association studies (GWAS) have linked many genetic loci with COPD and have contributed to the growing recognition that multiple biological mechanisms result in the observed lung function changes that receive the diagnostic label of COPD. Uncovering these so-called endotypes of COPD is a major goal of precision medicine that seeks to provide tailored medical care. However, research into sex-specific genetic risk factors that may inform sex-related COPD endotypes is lacking10.

The UK Biobank is a population-based cohort study with over 500,000 participants, which aims to investigate genetic and nongenetic factors that influence a variety of diseases affecting middle and older aged people11. Because the UK Biobank is the largest spirometric study ever conducted in the UK, it is an especially valuable resource to study COPD genetics compared to other biobanks. We previously showed that defining COPD according to Global Initiative for Obstructive Lung Disease (GOLD) criteria, which relies on lung function measures, yielded UK Biobank GWAS results consistent with those of genetic epidemiology cohorts but quite different from those based on ICD codes or self-reported disease12. Given its large sample size and the availability of many health-related measures and spirometry data for most genotyped participants regardless of health status, the UK Biobank provides an unprecedented opportunity to study sexually dimorphic features of COPD11,13.

Gene-based association methods are a complementary strategy for GWAS that evaluate biologically meaningful units of the genome14–16. They detect joint effects of multiple variants for which individual genetic effects may not reach genome-wide significance by reducing the burden of multiple tests that arise when simultaneously assessing millions of markers that are common to GWAS. In addition, gene-based statistics are often required as input for network- and pathway-based approaches. The versatile gene-based association study (VEGAS) method, which is commonly used to investigate putative genes for a trait of interest15,17, takes as input variant-level p-values and reference linkage
disequilibrium (LD) data and computes gene-based p-values using a simulation-based approach. Although VEGAS is offered as a webserver or local program that can be readily used by anyone with GWAS summary statistics regardless of the underlying study design, its use is hampered in practice when gene-based p-values are very small because the number of simulations required to obtain such p-values is very large. GWAS performed with biobanks that contain a large amount of clinical information linked to biological samples often yield sample sizes large enough to produce hundreds to thousands of nominal associations, which renders the VEGAS approach impractical due to the long run times necessary to obtain gene-based association results.

In this study, we identified sex-specific genetic associations of COPD using UK Biobank data, developed an R package that efficiently estimates gene-based p-values from GWAS summary statistics, and we compared SNP-level GWAS results with gene-based results obtained using the newly developed package.

**Methods**

**Study population and COPD definition**

Phenotypes of UK Biobank participants were obtained from responses to computer-assisted interviews, self-completed questionnaires, and physical measures obtained during in-person visits as reported in data-fields 31 (sex), 50 (height), 20003 (age), 20160 (ever smoked), 22001 (genetic sex) and 22006 (genetic ethnic grouping). To determine lung function, the ‘best’ measures per individual of FEV 1 and forced vital capacity (FVC) were obtained from pre-bronchodilator spirometry blow-volume time-series data following the spirometry quality control steps described in Shrine et al. COPD affection status was assigned based on spirometric evidence of moderate-to-severe airflow limitation by the modified GOLD criteria: cases had FEV 1/FVC < 0.7 and FEV 1 < 80% predicted while controls had FEV 1/FVC > 0.7 and FEV 1 > 80% predicted.

**Sex-stratified GWAS**

For each sex, a GWAS was performed using individuals who self-identified as white British and had very similar genetic ancestry based on genetic principal components, while excluding those who had a mismatch between self-reported and genetic sex as determined by chromosomal make-up, sex chromosome configurations that were not XX or XY, or had non-normal heterozygosity and missing rates according to genetic information provided by the UK Biobank team. These procedures resulted in 12,958 cases and 95,631 controls for males and 11,311 cases and 123,741 controls for females. At the genotype level, variants with imputation INFO score measure < 0.7 or minor allele frequency (MAF) < 0.01 were excluded. Genome-wide association testing was conducted using a generalized mixed model framework implemented in SAIGE (v0.43.3) to account for subject relatedness and fine-scale population structure, while including as covariates age, age-squared (age 2), height, ever-smoking status (ever vs never), and the first 4 principal components obtained from genotypes. Pack-years of smoking was not included due to a large amount of missing values as described in Shrine et al. FUMA, a web-based post-GWAS annotation tool, was used to identify independent risk loci meeting the genome-wide significance threshold of p-value = 5 × 10^-8 using the UK Biobank white British reference panel. Genetic correlation between the sex-specific GWAS results was estimated via LD score regression, while using pre-calculated LD scores based on the 1000 Genomes Project data for European populations. Regional association plots were created with LocusZoom (v1.4). SNP-by-sex interaction tests were performed for the sex-specific genome-wide significant loci using logistic regression models that included data for all participants and the same covariates used in the sex-specific analyses.

**Gene-based association tests**

For gene-based association tests, we employed the statistical model described in VEGAS. Briefly, the test statistic was defined as $Q = Z'Z = \sum_{i=1}^{p} z_i^2$, that is, the sum of squared variant-level z-statistics, where $Z = \{z_1, ..., z_p\}$ is a vector that follows a multivariate normal distribution with a mean vector $0$ and a covariance matrix $K$; i.e., $Z \sim N(0, K)$. VEGAS uses Monte Carlo simulations to approximate the distribution of $Q$, and thus, compute gene-based p-values. Instead of that approach, we rewrote $Q = XAX = \sum_{i=1}^{p} \lambda_i x_i^2$, where $K = \lambda \Sigma \chi^2$, (with a spectral theorem) and $X = P^\chi \Sigma^{-1/2} Z \sim N(0, I_p)$. This represents a quadratic form in independent central normal variables and its distribution can be evaluated with various methods such as numerical inversion of the characteristic function. We developed the R package `snpsettest` (https://CRAN.R-project.org/package=snpsettest) based on this latter version of the test statistic while utilizing the algorithm of Davies, or saddlepoint approximation, to obtain gene-based p-values. We performed gene-based association analyses with `snpsettest` using gene annotations from GENCODE release 19 for the sex-stratified COPD GWAS results. We considered genes in the following biotypes: protein-coding, immunoglobulin variable chain and T-cell receptor genes which included 18,774 entries. For each gene, any
variants within 20 kb of 5′ and 3′ UTRs were selected for association tests and the 1000 Genomes phase 3 European reference panel27 (genotypes of 240 males and 263 females) was used to infer relationships between markers. A significant association was defined as having a Bonferroni-corrected p-value < 0.05. Gene-based results were also obtained with a local instance of VEGAS217 using default parameters (up to one million simulations for p-value calculation) except for gene boundaries for SNP selection that were set to 20 kb of 5′ and 3′ UTRs of a gene.

Results

Sample characteristics

Characteristics of GWAS subjects stratified by sex are provided in Table 1. Male and female COPD cases were older and more likely to have ever smoked. The proportion of COPD cases was higher in males (11.9%) than females (8.4%), which is consistent with a higher portion of male subjects having a positive smoking history. Female subjects were more likely to have a diagnosis of asthma (ICD-based prevalence of 5.7% in male versus 7.2% in female subjects; self-reported doctor’s diagnosis prevalence of 10.9% in male versus 12.5% in female subjects), and a substantially greater proportion of COPD cases had asthma compared to controls. Consistent with the definition of COPD, lung function was lower in cases than controls as measured by percent predicted FEV1 and FEV1/FVC ratio.

Table 1. Characteristics of GWAS subjects

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Control</td>
<td>Case</td>
<td>Control</td>
</tr>
<tr>
<td>N</td>
<td>243,641</td>
<td>95,631</td>
<td>123,714</td>
</tr>
<tr>
<td>Age, mean (SD)</td>
<td>57.6 (8.4)</td>
<td>60.6 (7.7)</td>
<td>57.0 (8.3)</td>
</tr>
<tr>
<td>Male %</td>
<td>44.6</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Height (cm), mean (SD)</td>
<td>168.4 (9.1)</td>
<td>175.8 (6.9)</td>
<td>162.7 (6.3)</td>
</tr>
<tr>
<td>Ever smoker %</td>
<td>59.4</td>
<td>76.4</td>
<td>54.0</td>
</tr>
<tr>
<td>Asthma %</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ICD-coded diagnosis</td>
<td>9.7</td>
<td>19.3</td>
<td>5.8</td>
</tr>
<tr>
<td>Self-reported</td>
<td>11.8</td>
<td>10.6</td>
<td>32.9</td>
</tr>
<tr>
<td>Doctor’s diagnosis</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lung function, median (IQR)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FEV1 predicted percentage</td>
<td>96.3 (15.4)</td>
<td>67.5 (16.2)</td>
<td></td>
</tr>
<tr>
<td>FEV1/FVC ratio</td>
<td>0.77 (0.07)</td>
<td>0.78 (0.06)</td>
<td>0.78 (0.06)</td>
</tr>
</tbody>
</table>

ICD: International Classification of Diseases; IQR: Interquartile range.
* Affection status was assigned based on having any sub-categories of ICD-9 493 and/or ICD-10 J45.
† Affection status was assigned based on the UK Biobank data-fields 22127 (Doctor diagnosed asthma) and 20002 (Non-cancer illness code, self-reported).

Most COPD-associated loci were consistent by sex

We identified 17 and 14 genome-wide significant loci in male and female GWAS, respectively (Figure 1), each of which was previously identified in GWAS of COPD and/or lung function-related traits (e.g., FEV1, FEV1/FVC ratio, and smoking behavior). Nine of these loci were present in male and female GWAS results (Table 2), including the most well-replicated COPD locus near HHIP (Figure 2A). The estimated genetic correlation between male and female results was 0.961 (SE 0.04), demonstrating shared genetic effects overall. Male-specific associations were observed at loci in/near C1orf87, PEL11, ARHGEF3, C5orf56, PPP1R3B, BNC1, and CFDP1, and female-specific associations were observed at loci in/near MECOM, ADAM19, C10orf11, KANSL1, and SOGA2. Regional association plots for these genes suggested that some of the sex-specific associations were not due to differences in statistical power attributable to a decreased sample size, most notably that of the C5orf56 locus (Figure 2B). Results of the SNP-by-sex interaction analyses for lead variants are in Table 3. Consistent with the sex-stratified GWAS, rs2158101 at the male-specific locus in C5orf56 had the lowest interaction p-value (P = 1.16×10^-4). The other lead variants in the male-specific loci had interaction p-values < 0.05 except for rs330934 at the PPP1R3B locus. Among the lead variants in female-specific loci, only rs2637261 at the C10orf11 locus had an interaction p-value < 0.05.

The snpsettest package results were comparable to those of VEGAS

Comparison of gene-based association tests obtained with snpsettest versus those of VEGAS found a strong agreement between the two: estimated correlations of -log10(p-value) were 0.981 for the male-specific GWAS and 0.991 for the female-specific GWAS (Figure 3). We observed a lack of consistency with VEGAS for genes whose actual p-values
were < 10^{-6}. When only the gene-based p-values ≥ 10^{-6} were considered, the correlations were 0.999 for both sets of results. A major advantage of the \textit{snpsettest} package over VEGAS was speed. On the same laptop (i9-9980HK, 32GB memory, Windows subsystem for Ubuntu 18.04 LTS), VEGAS took 29 and 32 hours to complete the gene-based association tests for 18,774 genes using ~8.8 million GWAS summary statistics of males and females, respectively, while the \textit{snpsettest} package completed the tests in approximately 9 minutes.

**Figure 1.** Manhattan plots of sex-stratified COPD GWAS results for (A) male and (B) female subjects. Each locus was annotated with its nearest coding gene. The green horizontal dashed line indicates the genome-wide significance threshold of 5 \times 10^{-8}. Overlapping loci in the GWAS are shown in blue. Sex-specific loci are shown in orange when genome-wide significant, with corresponding non-significant region shown in red in the alternate sex-specific plot.

**Sex-specific gene-based association tests identified two additional loci of interest**

Sex-specific gene-based associations were determined as those meeting a Bonferroni-adjusted threshold, while excluding loci in the HLA region (Figure 4). Consistent with the GWAS results, \textit{C5orf56} on 5q31.1 had the strongest sex-specific association. Other male-specific associations included \textit{OTUD4}, \textit{ABCE1}, \textit{ANAPC10} on 4q31.21; \textit{FBXO38} on 5q32; and \textit{TMEM170A}, \textit{CFDP1}, and \textit{CHST6} on 16q23.1. Female-specific associations were \textit{ASTN2} and \textit{TRIM32} on 9q33.1; \textit{LCAT}, \textit{CTRL}, and \textit{PSMB10} on 16q22.1; and \textit{NSF}, \textit{CRHR1}, and \textit{SPPL2C} on 17q21.31. Most of the significant genes were in/near the genome-wide significant GWAS loci, but novel sex-specific gene associations on 16q22.1 and 9q33.1 were observed only via gene-level statistics.

**Discussion**

COPD is a complex disease with sex-specific differences in susceptibility and presentation\textsuperscript{10}. Previous studies by COPDGene (Genetic epidemiology of COPD) found that sex-related genetic components conferred a higher risk of severe, early-onset COPD in women\textsuperscript{33,34}, but the genetic contributions that lead to sex divergence in COPD remain poorly understood. In sex-stratified COPD GWAS, we identified 17 genome-wide significant loci for males and 14 for females. Although the number of significant loci identified was smaller than that in our previous study\textsuperscript{12} that used all subjects—likely due to decreased statistical power—all loci have been previously associated with COPD and/or lung function-related traits\textsuperscript{21,35}. Our results demonstrated that a large proportion of genetic liability are shared between males and females: i) there were 9 overlapping genome-wide significant loci corresponding to genes \textit{FAM13}, \textit{NPNT}, \textit{HHIP}, \textit{HTR4}, \textit{HLA-DQB1} (\textit{HLA-DQA1}), \textit{GRP126}, \textit{CDC123}, \textit{THSD4}, and \textit{CHRNA5} (\textit{HYKK}), and ii) there was substantial genetic correlation when considering the effects of all variants not reaching genome-wide significance.

Sex-specific genome-wide significant associations were observed in 8 male-specific and 5 female-specific loci. SNP-by-sex interaction tests for these loci using data for all subjects supported most of the stratified findings except for the male-specific locus in \textit{PPP1R3B} and female-specific loci in/near \textit{MECOM}, \textit{ADAM19}, \textit{KANSL1}, and \textit{SOGA2}. The
locus in C5orf56 showed the largest difference in patterns of association: it was convincingly associated with COPD in males but no signal was present in females. C5orf56 is a long non-coding gene known as IRF1-ASI that has not been mechanistically linked to COPD, but the nearby IRF1 gene encodes interferon regulatory factor 1, which has been associated with anti-viral defense in airway epithelium\(^6\). Given that respiratory viral infections are a common cause of exacerbations of chronic lung diseases and that viruses (e.g., Human rhinovirus, respiratory syncytial virus, and influenza) are often detected during COPD exacerbations specifically\(^37,38\), the associated locus may influence susceptibility to COPD via altered responses to virus exposure. Interestingly, the IRF1 locus has been proposed as a strong candidate region for male-specific asthma susceptibility that may attributable to sex-specific interferon responses\(^39\). Another male-specific significant locus we observed near PEL1 may be linked with response to airways viruses\(^40\) given that PEL1 is involved in IL-1 signaling and its expression was correlated with the number of exacerbations experienced by patients with obstructive airway disease\(^41\). Other regions with sex-specific associations in ARHGEF3, C1orf87, and C10orf11 are not easily linked to mechanistic hypotheses based on what is currently known about the function of these genes.

### Table 2. Summary statistics for the lead variants at genome-wide significant loci

<table>
<thead>
<tr>
<th>rsID</th>
<th>Position†</th>
<th>RA</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Male RA</td>
<td>Female OR</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>OR</td>
<td>P-value</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Male P-value</td>
<td>Female P-value</td>
</tr>
</tbody>
</table>

**Male risk loci**

- rs1420472  3:168776326 T 0.44 1.06 (1.03-1.10) 9.69×10⁻⁴ 0.44 1.00 (0.96-1.04) 1.14×10⁻⁴
- rs2045517† 4:89070964 T 0.41 1.04 (1.01-1.07) 8.68×10⁻⁴ 0.41 1.01 (0.97-1.06) 9.12×10⁻⁴
- rs3172297† 4:15492003 C 0.40 0.83 (0.80-0.86) 5.73×10⁻⁴ 0.40 0.83 (0.80-0.86) 3.69×10⁻⁴
- rs5866500† 5:12756232 C 0.43 0.86 (0.84-0.89) 1.48×10⁻⁸ 0.43 0.86 (0.84-0.89) 1.26×10⁻⁹

**Female risk loci**

- rs1843712 5:18591385 AGCCGG 0.34 1.07 (1.04-1.10) 6.66×10⁻⁸ 0.34 1.07 (1.04-1.10) 7.97×10⁻⁸
- rs17843068 6:32620545 C 0.48 1.09 (1.06-1.13) 6.30×10⁻⁴ 0.48 1.12 (1.09-1.16) 5.59×10⁻⁴
- rs26115† 6:142817407 C 0.31 0.84 (0.80-0.88) 7.36×10⁻⁷ 0.31 0.87 (0.82-0.92) 1.06×10⁻⁷
- rs2001546† 10:12285735 T 0.52 0.92 (0.89-0.95) 3.60×10⁻⁴ 0.51 0.92 (0.89-0.95) 7.44×10⁻⁴

RA: Risk allele; RAF: Risk allele frequency; OR: Odds ratio

* Position based on GRCh37, † Overlapping loci according to genomic position

To obtain gene-based associations from SNP-level results, we developed the *snpsettest* R package that modifies the statistical test in the VEGAS software\(^15,17\) to compute gene-based p-values more efficiently. Although VEGAS can run in a reasonable amount of time if the maximum number of simulations is bounded (e.g., 10⁴), this truncation is not optimal when gene-based association statistics are needed for large-scale GWAS with many significant SNP
associations, and subsequently, many gene-based associations that are necessary as input for subsequent pathway and network analyses. We demonstrated that our package produces results consistent with those of VEGAS but with a much shorter runtime.

Figure 2. Regional association plots for the loci in (A) HHIP as an example of a finding present in male and female GWAS and (B) C5orf56, the locus with the greatest sex-specific difference. Left panels display the results of the male GWAS and the right panels display the results of the female GWAS.

By comparing gene-level associations obtained for sex-specific COPD GWAS, we found that C5orf56 had the strongest sex-specific COPD association, as it was present only in males. Other male-specific associations were found in CFDP1, TMEM170A, and CHST6 located in the non-overlapping significant GWAS locus on 16q23.1. This region has been implicated in coronary heart disease, whose mortality is higher among men\textsuperscript{42,43}, and was identified as having a potential genetic overlap with COPD susceptibility\textsuperscript{35}. Although OTUD4, ABCE1, and ANAPC10 on 4q31.21 and FBXO38 on 5q32 had male-specific associations, they had similar patterns of regional variant associations despite not reaching genome-wide significance levels in the female GWAS. The most prominent female-specific associations were observed in ASTN2 and TRIM32 on 9q33.1, a region previously reported as having sex-specific associations with neurodevelopmental disorders\textsuperscript{44}. Moreover, several genes on 16q22.1 were associated with COPD only in females, of which SMPD3 was previously

<table>
<thead>
<tr>
<th>rsID</th>
<th>Position</th>
<th>Risk Allele</th>
<th>Beta</th>
<th>P-value*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male-specific risk loci</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs77714938</td>
<td>1:60926112</td>
<td>C</td>
<td>0.105</td>
<td>2.19×10^{-2}</td>
</tr>
<tr>
<td>rs572473905</td>
<td>2:64288008</td>
<td>T</td>
<td>-0.283</td>
<td>2.84×10^{-4}</td>
</tr>
<tr>
<td>3:5706052_GCCCA_G</td>
<td>3:5706052</td>
<td>G</td>
<td>0.095</td>
<td>5.63×10^{-4}</td>
</tr>
<tr>
<td>rs2158101</td>
<td>5:131769273</td>
<td>A</td>
<td>0.091</td>
<td>1.16×10^{-4}</td>
</tr>
<tr>
<td>rs330934</td>
<td>8:9013766</td>
<td>T</td>
<td>0.16</td>
<td>0.441</td>
</tr>
<tr>
<td>rs7433294</td>
<td>9:101673848</td>
<td>G</td>
<td>-0.056</td>
<td>1.07×10^{-2}</td>
</tr>
<tr>
<td>15:84034556_CA_C</td>
<td>15:84034556</td>
<td>C</td>
<td>-0.082</td>
<td>3.69×10^{-4}</td>
</tr>
<tr>
<td>rs72787160</td>
<td>16:754545162</td>
<td>C</td>
<td>0.048</td>
<td>1.76×10^{-2}</td>
</tr>
</tbody>
</table>

| Female-specific risk loci   |               |             |        |          |
| rs1120472                   | 3:168776326   | T           | -0.027 | 0.179    |
| rs112325689                 | 5:156931851   | AGCCGG      | -0.038 | 6.86×10^{-2} |
| rs2637261                   | 10:78320593   | T           | -0.048 | 1.73×10^{-2} |
| rs113434679                 | 17:44126765   | A           | -0.039 | 0.121    |
| 18:8803157_TA_T             | 18:8803157    | T           | 0.036  | 0.105    |

*P-values from logistic regressions with female as a reference group.
reported as having gene-by-smoking interaction effects on COPD\textsuperscript{45}, suggesting that the 16q22.1 locus may be involved in sex-related differences in lung function decline between male and female smokers.

Our study is limited in that some of the sex-specific associations may not be observed due to insufficient statistical power resulting from decreased sample sizes. We chose to perform stratified analyses for ease of interpretation of association odds ratios, and we note that identification of interactions in general requires more statistical power than that of a standard GWAS because effect sizes of interactions are expected to be smaller than their main effects. Future replication of results in independent cohorts could ensure generalizability of our findings. Our study lacks functional validation of sex-specific genetic components in support of the statistical associations, but future experimental studies can explore our findings to elucidate the mechanisms underlying the observed differences. Another limitation is that the UK Biobank is not representative of the general population with respect to a variety of health characteristics; participants were more likely to be older, to be female, and to be more health-conscious than nonparticipants\textsuperscript{46}. Additionally, our results were obtained from participants of European ancestry, and thus, may not generalize to other racial/ethnic groups. Our gene-based association tests did not fully utilize information from the GWAS results since the lead variants in some loci were not necessarily within boundaries of protein-coding genes.

In summary, sex-stratified GWAS of COPD found substantial overlap in the significant risk loci and genetic correlation of male versus female results, but evidence of sex-specific effects was found for several genes, the most prominent of which were \textit{C5orf56, CFDP1, TMEM170A, CHST6, ASTN2} and \textit{TRIM32}. We developed the \textit{snpsettest} package to conduct gene-based association tests with GWAS summary statistics and identified genes showing a sex-specific association. Contrasting the GWAS and gene-based association test results provided insight into what genetic components can be the topic of future studies aimed at understanding sexual dimorphisms in COPD.

\textbf{Acknowledgements}

This work was supported by National Institutes of Health (NIH) awards R01 HL133433 and R01 HL141991. This research was conducted using the UK Biobank Resource under Application Number 40375.

\textbf{References}


8. Gan WQ, Man SFP, Postma DS, Camp P, Sin DD. Female smokers beyond the perimenopausal period are at increased risk of chronic obstructive pulmonary disease: a systematic review and meta-analysis. Respir Res. 2006 Mar 29;7:52.


Automated Mapping of Real-world Oncology Laboratory Data to LOINC

Jonathan Kelly, MEng¹, Chen Wang, MSHI², Jianyi Zhang, MSHI², Spandan Das, MEng¹, Anna Ren, BSE¹, Pradnya Warnekar, MSHI¹
¹Flatiron Health Inc, New York, New York; ²Georgetown University, Washington D.C.

Abstract
In this study we seek to determine the efficacy of using automated mapping methods to reduce the manual mapping burden of laboratory data to LOINC® on a nationwide electronic health record derived oncology specific dataset. We developed novel encoding methodologies to vectorize free text lab data, and evaluated logistic regression, random forest, and knn machine learning classifiers. All machine learning models did significantly better than deterministic baseline algorithms. The best classifiers were random forest and were able to predict the correct LOINC code 94.5% of the time. Ensemble classifiers further increased accuracy, with the best ensemble classifier predicting the same code 80.5% of the time with an accuracy of 99%. We conclude that by using an automated laboratory mapping model we can both reduce manual mapping time, and increase quality of mappings, suggesting automated mapping is a viable tool in a real-world oncology dataset.

Background and Significance
Health data collected in the course of routine clinical care (real-world data [RWD]) are becoming a valuable part of the clinical research armamentarium, complementing and/or supplementing traditional prospective studies, and providing insights on aspects such as patterns of care or treatment effectiveness in populations underrepresented in clinical trials.¹,² Electronic health records (EHRs) have emerged as a key oncology RWD source, with the potential to generate highly granular longitudinal data.³ The original purpose of EHRs however, is not research but patient care, administration, and reimbursement. Therefore, extraction of research-grade information from the EHR may become a multi-step process that requires optimization and quality controls.⁴

The utilization of EHR-derived data for research purposes involves extraction of structured and coded data, as well as unstructured data (narrative free text entered by users at the point of care). One domain in particular that requires significant pre-processing is that of laboratory (lab) data.⁵,⁶ Laboratory data is crucial in describing the longitudinal patient journey and thus high quality laboratory data is essential in many areas of oncological research. For example, laboratory data is necessary to understand the efficacy of a patient’s regimen or the performance of novel treatments.⁷ However inconsistent use of standards by laboratories and free text documentation leads to highly variable lab results data.⁶,⁸ The Logical Observation Identifiers Names and Codes (LOINC®)⁹ is a vocabulary standard used to identify and unify lab data under a common data model, but is often inappropriately or inconsistently used in clinical day-to-day settings.¹⁰,¹¹ Due to this, a manual mapping process where specialists assign laboratory data to LOINC codes is required before the data can be used in research datasets.¹²

This manual mapping process, however has some shortcomings:

1. Ingestion of standardized and non-standardized lab data from multiple vendors, free text data entry at practice sites, and abbreviations, typos, and practice-specific documentation norms lead to a potentially infinite set of source terms.⁵,⁸ In our study we have found that it takes an experienced clinical terminologist between 6-8 hours to map 1000 terms manually, meaning harmonizing large amounts of laboratory data to LOINC codes is an arduous process.

2. As with any manual process, there is inevitable error in the mapping process, in spite of having mapping guidelines. As any error affects the accuracy of data that will be used for research and analysis, multiple levels of mapping review are required, further increasing the manual workload.

3. Any free text terms that come in from source data not previously harmonized are required to be mapped before they are eligible for entry into research datasets. Depending on volume, harmonization can be a lengthy process, reducing the data recency of datasets.
Based on the above considerations, finding an automated approach to lower the burden and increase the accuracy of the mapping process has direct implications on the quality of real-world research datasets. As a single mapping can affect many rows in a database, high accuracy mappings are of top importance when examining automated approaches.

Previous studies have attempted to automate LOINC mapping in a number of different scenarios. One study attempted to use a local high quality corpus and was able to achieve a best case accuracy of 79%. A second study that relied on using lexical methods achieved similar accuracy. A more recent study trained a machine learning classifier on a large national EHR database with noisy LOINC labels, and was able to achieve relatively high accuracy. The best classifiers in this study predicted the correct LOINC code in 85% of the unlabeled data and 96% of the labeled data by test frequency. More recently in a study focused on mapping COVID-19 labs to LOINC codes, a rules based algorithm was shown to have accuracy of 97.4%, prompting further investigation into deterministic rules based algorithms. Other studies have been successful at higher level grouping of laboratory data into categorical values, but did not focus on the standardization of individual lab records.

To the best of our knowledge, there are no studies that have attempted to use automated LOINC mapping on a real-world dataset actively being used for clinical research, or on an oncology-specific dataset. These properties lead us to develop custom encoding methodologies, with the goal of high accuracy and high precision automated LOINC mappings. We used these novel encoding methodologies to process free text laboratory data for the use in supervised machine learning classifiers. We evaluated the ability for these classifiers to individually predict LOINC codes, as well as the ability for groups of classifiers to jointly make predictions through ensemble learning. Any reduction in manual mapping time or manual quality assessment work that an automated system can provide to clinical terminologists (without compromising accuracy) is our overall measure of success.

**Methods**

**Index Of Lab Data**

This study used the nationwide, longitudinal Flatiron Health electronic health record (EHR)-derived de-identified database. During the study period, the de-identified data originated from approximately 280 US cancer clinics (~800 sites of care). Flatiron Health creates EHR-derived research datasets comprising de-identified patient-level structured and unstructured data, curated via technology-enabled abstraction. Flatiron Health has a harmonization process whereby research relevant clinical and administrative data are mapped by clinical terminologists to their appropriate terminology standards. All free text laboratory data undergoes this process, getting assigned LOINC codes. Once data is harmonized, it is re-used to determine standard codes for any current and future free text data.

Flatiron Health lab data is manually harmonized using term, unit, and panel information, with each distinct combination corresponding to a different row that each requires harmonization. As all three fields can potentially be free text fields in EHRs, thousands of new combinations require mapping each month.

The “term” field of the lab data contains the name of the lab result. This can vary from a fully spelled out name to a local acronym. An example of the term field is “white blood cell count”, which also might appear as “white blood cell”, “wbc”, “wite bld cell”, or any other abbreviation with or without typos.

The “unit” field aptly contains information about the unit of measurement. Examples of a few possible unit values can be seen in Table 1. As shown in Table 1, rows that have the same term value but differing unit values can result in different target LOINC codes. Lab results can also be expressed in equivalent units, for example the units mg/mL, g/dL and ug/mL, all measure mass/volume, but create separate rows for mapping.

The “panel” field represents a group of lab tests/results that are ordered and reported together. There are several bases for grouping individual labs together. For example, by the medical condition they are intended to help diagnose (cardiac risk panel), by the specimen type (complete blood count, CBC), by the tests most frequently requested by users (comprehensive chemistry profile), by the methodology employed in the test (viral panel by polymerase chain reaction), or by the types of components included (urine drug screen). Since the specimen for the lab result is not available as a distinct data point, we use the panel name to determine the system. In the Flatiron Health dataset the panel field contains the concatenation of all the panels that were ordered, adding variance and further increasing the amount of data that requires harmonization.
Datasets

There are three catalogs we used in our experiments for automated lab mapping harmonization:

1. A source data catalog
2. A target LOINC data catalog
3. A mapping catalog containing associations between the source and target catalogs

The source data catalog contains all of the unique term, unit, and panel combinations that we have ever seen in our data, as well as an associated source ID. This table contains no LOINC code information. The “target” LOINC catalog contains all of the different LOINC codes that we might map the rows in the source data catalog to. Our harmonization team takes in the source catalog as input and manually determines which target code each row in the source catalog should map to.

Once a row from the source data catalog has been assigned an appropriate LOINC, it is added to the mapping catalog. Thus, this catalog contains the source ID and the associated target ID (LOINC) in a unique mapping. Despite the full LOINC catalog containing approximately 95,000 distinct LOINCs, because the source dataset is oncology specific we see only a small percentage of all possible LOINCs in our dataset. If the harmonization team determines that the information in the source row is insufficient to determine an accurate mapping, a label of “EXCLUDED” is attached and no target code is present. This provides an additional challenge for an automated system, as this label has a wide range of source data mapped to it, making it challenging to accurately predict.

There are different levels of LOINC interoperability, depending on the differences between two LOINC codes. Previous work has consolidated LOINC codes using these levels of interoperability\textsuperscript{17}, but since the Flatiron Health dataset requires the most granular version of LOINCs, we do not do consolidation. While this makes automated LOINC prediction more challenging, it is required to keep the dataset as precise as possible.

Deterministic Automated LOINC Prediction

Before using learning-based automated LOINC prediction algorithms, we first set baselines using two computation-based algorithms. The two algorithms, maximal target string matching and maximal source string matching, are described below.

Maximal Target String Match

The first baseline algorithm we used compared source rows directly to the target LOINC catalog. We compared using Levenshtein distance, which is a string metric for measuring the difference between two character sequences.\textsuperscript{22} Specifically we used the Levenshtein ratio, which is a number between 0 and 1 representing how similar two strings are, with 0 being completely different and 1 being entirely the same. For each row in the source catalog, we compared the source term and source unit fields directly to the target LOINC catalog. We then combined the comparisons via a weighted sum of term and unit. This set of weights was evaluated against six other combinations ($\frac{4}{5}$ term and $\frac{1}{5}$ unit, $\frac{3}{5}$ term and $\frac{2}{5}$ unit, $\frac{2}{5}$ term and $\frac{3}{5}$ unit, $\frac{1}{5}$ term and $\frac{4}{5}$ unit, $\frac{2}{5}$ term and $\frac{4}{5}$ unit, $\frac{4}{5}$ term and $\frac{2}{5}$ unit) on a random sample of 100,000 rows from the source term catalog and had the highest performance of the six. Since the panel field contains a combination of all panels that were ordered, it has no directly comparable value in the LOINC catalog and was therefore excluded. For each row, the LOINC code
that was most similar via the weighted sum of term and unit Levenshtein ratios was the one assigned.

Maximal Source String Match
The second baseline algorithm attempts to find the source row in the training data that is most similar to the test row being evaluated. Similar to the Maximal Target String Match algorithm, it uses Levenshtein ratio to measure similarity. To compare two rows, the Levenshtein ratio of the term, unit, and panel fields are calculated and combined via a weighted sum of \( \frac{3}{5} \) term, \( \frac{1}{5} \) unit, and \( \frac{1}{5} \) panel. We again tried multiple weighted combinations of fifths, and this one had the highest performance of the combinations we evaluated. Computing the Levenshtein ratio is a costly operation, and so comparing each row in the training data to every row in the source data is prohibitively expensive. Due to this limitation, we instead first pre-compute the five most common source terms for each LOINC, with ties being broken by random selection. We then compare each row in the test data to all of the pre-chosen source rows, and choose the LOINC that has a source row with the greatest similarity.

Feature Encodings
As the source data is composed of free text fields, we must first encode the free text to make it possible to pass as an input to a supervised machine learning model. Since lab data free text fields have an unlimited set of potential values, and are often short abbreviations, we found that traditional string encoding methods that rely on a finite corpus to encode such as One-Hot tokenization, TF-IDF, and Word2Vec perform poorly.23–25 With this in mind, we developed two new encoding methods specifically for this domain.

Levenshtein Distance Encoding
As mentioned above, Levenshtein distance is a string metric for measuring the difference between two character sequences.22 To encode each row, we compared the rows source term and source unit to every target description and target units in the LOINC target catalog, for each distinct LOINC in the training data. The weighted sum of the Levenshtein ratio for term and unit between the source catalog and each row in the LOINC catalog is captured in a vector. To determine the optimal set of weights to use, we tried every tenth fold combination (\( \frac{1}{10} \) term, \( \frac{1}{10} \) unit, ..., \( \frac{9}{10} \) term and \( \frac{1}{10} \) unit) and settled on \( \frac{8}{10} \) term and \( \frac{2}{10} \) unit. The vector of weighted Levenshtein ratios, which has a column for each LOINC representing how similar the source row is to that target LOINC, is the encoded row. Similar to the Maximal Target String Match, we exclude using panel as it has no directly comparable value in the LOINC catalog to compare to with Levenshtein distance.

Frequency Tokenization Encoding
In a similar manner to TF-IDF, we created an approach to encode based on the frequency of distinct tokens in the source data. For each row in the training data set, we first clean the rows by removing any characters that are not alpha-numeric and replacing them with whitespace. This process is done for each of the term, unit, and panel fields. We then split the cleaned row on whitespace, creating a set of tokens for each field. These tokens are then mapped to the target LOINC, creating a token-LOINC map that details which source data tokens are associated with each LOINC. If tokens appear multiple times mapping to the same target LOINC, that count will be recorded in the token-LOINC map. This token-LOINC map can then be used to encode future rows.

To encode the term field, we first create tokens in the same manner as described above. These tokens will be used to create an encoded vector of length L, where L is the total number of LOINCs in the token-LOINC map. In this vector, each LOINC has a corresponding index, initialized at zero. For each LOINC that a token maps to, the corresponding index in the vector will be incremented by the count in the token-LOINC map. See Figure 1 for an example of this encoding. This process happens for each of the term, unit, and panel fields, resulting in a concatenated encoded vector of length 3*L.

Supervised Machine Learning Classification Models
For both of the feature encoding methods, we evaluate performance using Logistic Regression (L2 penalized)26, Random Forest27, and K Nearest Neighbors (KNN)28 classifiers. Model training and analyses were conducted using scikit-learn in Python 3.7.4,29
Figure 1: Tokenization Encoding process. Training data gets processed and tokenized, forming the token-LOINC map. This map is then used to encode both the training and test data, leading to vectorized rows. The encoded row has an index for each LOINC that exists in the training data catalog (here only 2 LOINCs), and the value at each index represents how many tokens overlap with training data tokens mapped to that LOINC. The first row gets encoded as \([3, 0]\) since there are 3 tokens in the source term that match a token for the LOINC 26464-8, and 0 tokens that match the LOINC 26499-4.

We used GridSearch to optimize key hyperparameters for each model. For logistic regression we focused on the number of iterations, and the inverse regularization parameter (C), and settled on values of 500 and 5 respectively. For random forest we focused on the number of estimators, and settled on 500. For KNN we focused on the number of nearest neighbors, and settled on 5.

Additionally, we evaluated the use of PCA to reduce the size of the input vectors and thereby decrease learning time. Using PCA with 95% variance we were able to decrease the total learning time by 50%. However, the increased efficiency came at the cost of a few percentage points of accuracy. As high accuracy is our top priority, the results in this paper do not use PCA or other compression schemes on the input vectors.

Ensemble Learning

Ensemble learning is the practice of using multiple learning algorithms together and then classifying new points based on a weighted vote of their individual predictions. We explore the use of ensemble learning to increase accuracy of predictions in cases where multiple models predict the same LOINC. Our ensemble models are assessed based on the percent of the test set with prediction overlap, and the accuracy of the predictions on the overlap. The percent of the test set with overlap is simply the number of rows in the test set where all models make the same prediction, divided by the size of the test set. The accuracy is then measured in the standard way on the subset: by taking the predictions that all models made and comparing with the labeled data. A high performing ensemble model will have both a high percentage of test set overlap, and high accuracy on the overlap.
It is important to note that adding additional models to an ensemble with these assessment criteria will not necessarily increase the ensemble’s performance. For example, because we define prediction overlap to be only predictions where all models make the same prediction, adding a poorly performing model may significantly reduce prediction overlap, decreasing the subset the ensemble is able to make a prediction on. Moreover, a poorly performing model may only overlap on incorrect predictions, significantly decreasing the ensemble’s accuracy. With this in mind, we evaluate various ensembles to attempt to find the set of models that most successfully complement each other.

Data Size and Model Performance

Models were trained using a subset of 450,000 randomly sampled and unique rows from the most recent two years of the labeled data corpus. We decided to use a large sample instead of training on the full dataset to increase investigative agility while maintaining high usability. Amongst the rows sampled, there are 482 distinct LOINCs that have been mapped to, including “EXCLUDED”. Models were validated using an 80/20 split for training and testing respectively, and results are the average of 5 randomly subsampled splits. Models were assessed based on accuracy, F-1 score, and precision. Since models are multiclass classifiers, F-1 score and precision are measured as weighted averages of the F-1 score and precision across all classes. While accuracy is of the highest concern for prediction usability, F-1 score is important to assess whether accuracy is overfitting due to class imbalance and thereby simply predicting the most frequent LOINCs.

Results

Deterministic Algorithm Performance

To establish a baseline we first built and tested the deterministic prediction algorithms. These algorithms do not take advantage of larger data sets to learn, so modifying the size of the dataset only affects the variance. With this in mind, we averaged predictions from 5 trials run using 10,000 randomly sampled rows. The performance of these algorithms is in Table 2. For Maximal Target String Match, accuracy is low, but still much higher than random chance, implying that there is some inherent similarity between the source data and the target data. Given the level of abbreviation in the source data, this result is expected. Comparing to other source data instead of directly to target LOINC data in the Maximal Source String Match algorithm, performed significantly better. When solely basing future predictions on the most similar previously seen terms, we saw an accuracy of just below 50%. Incorporating unit and panel information as well, we saw a small increase, implying that there is value in using more than just the lab name represented in the term field when making predictions.

<table>
<thead>
<tr>
<th>Algorithm Name</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximal Target String Match (Term + Unit)</td>
<td>16.0</td>
</tr>
<tr>
<td>Maximal Source String Match (Term)</td>
<td>48.4</td>
</tr>
<tr>
<td>Maximal Source String Match (Term + Unit + Panel)</td>
<td>52.0</td>
</tr>
</tbody>
</table>

Table 2: Deterministic (non-learning) algorithms performance for predicting LOINC, run on 10,000 LOINCs.

Cross Validated Supervised Model Performance

After determining a baseline, we moved on to assessing the performance of different learning based models using our two different encoding schemes. The results from the Levenshtein Distance Encoding method are in Table 3. We can see that we have significantly outperformed the baseline, and that while the random forest classifier had the highest accuracy, all models have comparable performance at greater than 93% accuracy. KNN has marginally worse performance using this encoding scheme, but it is worth noting that training time for the KNN classification model is on average less than 50% that of logistic regression or random forest. We also see that the weighted F1 score and weighted precision have similar values to that of accuracy, implying that our models are not overfitting due to class imbalance.

As we are focused on any reduction in manual work, it is also important to see if there exist particular subsets that our
models can reproducibly predict with higher success. One such subset we examined is the top 10% of LOINCs by labeled rows, a subset which on average accounts for over 84% of the test set. Unsurprisingly our models have a higher weighted precision on LOINCs in the top 10%, as these LOINCs have significantly more training data. However they still do relatively well on LOINCs that have less training data, once again suggesting that overfitting is not occurring.

Examining the results from the Frequency Tokenization Encoding in Table 4, we see that only the random forest classifier is able to outperform any model using the Levenshtein distance encoding. We also see that in this encoding method, random forest outperforms both logistic regression and KNN by a significant margin. Interestingly, when using the deterministic baseline algorithms, comparison to the LOINC catalog did significantly worse than comparison to the source data directly. Contrastly, when using supervised learning the Levenshtein encoding which encodes by comparing directly to the target LOINC catalog on average outperforms token encoding which encodes by comparing to previously seen source data. We also see slightly more variance between accuracy and weighted F1 Score as well as between the top 10% and bottom 90% of LOINCs, implying that there might be more class imbalance overfitting happening.

<table>
<thead>
<tr>
<th></th>
<th>Accuracy (%)</th>
<th>F1 Score (Weighted)</th>
<th>Precision (Weighted)</th>
<th>Top 10% LOINC Weighted Precision n=74866</th>
<th>Bottom 90% LOINC Weighted Precision n=15134</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic Regression</td>
<td>93.1</td>
<td>0.929</td>
<td>0.928</td>
<td>0.941</td>
<td>0.874</td>
</tr>
<tr>
<td>Random Forest</td>
<td>94.0</td>
<td>0.940</td>
<td>0.938</td>
<td>0.949</td>
<td>0.899</td>
</tr>
<tr>
<td>KNN</td>
<td>93.0</td>
<td>0.931</td>
<td>0.928</td>
<td>0.946</td>
<td>0.854</td>
</tr>
</tbody>
</table>

Table 3: Model performance using Levenshtein Distance Encoding trained using 80/20 split on dataset of size 450,000 (360,000 training, 90,000 testing). Top 10% and bottom 90% are measures of target LOINC frequency in testing dataset.

Table 4: Model performance using Frequency Tokenization Encoding trained using 80/20 split on dataset of size 450,000 (360,000 training, 90,000 testing). Top 10% and bottom 90% are measures of target LOINC frequency in testing dataset.

**Ensemble Learning Performance**

While supervised learning models did significantly better than baseline models, accuracy still has room for improvement. To increase usability, we attempted to use ensemble learning to see if we can achieve a higher accuracy for a subset of the predicted data. Results for different groupings of trained models are in Table 5. These results are extremely promising, with the combination of all trained models achieving an extremely high accuracy of 99% on the 80.5% of predictions that all models agreed upon. While using this method doesn’t allow for full automation as it is only applicable to a subset of predictions, it has the potential to significantly reduce the work of manual harmonization and quality analysis.

In these experiments we observe that the combination of models across encoding methods generally leads to better performance when compared to the combination of models within an encoding method. One potential reason for this is that a feature of the encoding method itself leads different models to make similar predictions. This would also explain the high prediction overlap of the models trained with the Levenshtein distance encoding. As the different encoding methods emphasize unique aspects of the source data, the combination of the encoding methods allows the
Table 5: Ensemble learning performance for various combinations of models and encoding methods.

<table>
<thead>
<tr>
<th>Model Combination</th>
<th>Prediction overlap raw count (out of 90000)</th>
<th>Percentage of Test Set (%)</th>
<th>Accuracy on prediction overlap (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR + KNN (Levi)</td>
<td>86430</td>
<td>96.0</td>
<td>95.2</td>
</tr>
<tr>
<td>LR + RF (Levi)</td>
<td>87580</td>
<td>97.3</td>
<td>94.9</td>
</tr>
<tr>
<td>KNN + RF (Levi)</td>
<td>87211</td>
<td>96.9</td>
<td>95.1</td>
</tr>
<tr>
<td>LR + KNN (Tokens)</td>
<td>78469</td>
<td>87.2</td>
<td>94.7</td>
</tr>
<tr>
<td>LR + RF (Tokens)</td>
<td>79174</td>
<td>88.0</td>
<td>97.1</td>
</tr>
<tr>
<td>KNN + RF (Tokens)</td>
<td>81643</td>
<td>90.7</td>
<td>96.3</td>
</tr>
<tr>
<td>LR Levi + Tokens</td>
<td>80003</td>
<td>88.9</td>
<td>95.9</td>
</tr>
<tr>
<td>RF Levi + Tokens</td>
<td>83526</td>
<td>92.8</td>
<td>97.9</td>
</tr>
<tr>
<td>KNN Levi + Tokens</td>
<td>79103</td>
<td>87.9</td>
<td>97.0</td>
</tr>
<tr>
<td>LR + RF + KNN (Levi)</td>
<td>85774</td>
<td>95.3</td>
<td>95.6</td>
</tr>
<tr>
<td>LR + RF + KNN (Tokens)</td>
<td>75423</td>
<td>83.8</td>
<td>97.7</td>
</tr>
<tr>
<td>LR + RF + KNN (Levi + Tokens)</td>
<td>72477</td>
<td>80.5</td>
<td>99.0</td>
</tr>
</tbody>
</table>


models to escape issues inherent to the encoding, generating a more robust prediction. We also observe that within encoding methods, increasing the number of models causes a drop in the overlap percentage, but results in a higher accuracy in the overlap.

Discussion

This study has shown that automated machine learning methods can be successful in mapping laboratory data to LOINC codes in a real-world oncology dataset. Overall our best performing model had an accuracy of 94.5% on the full hold-out dataset, and our ensemble method had an accuracy of 99.0% on 80.5% of the hold-out dataset. Additionally we have shown that it’s possible to make predictions to LOINCs with the same level of specificity as clinical terminologists, including marking rows with insufficient information as EXCLUDED. Success in mapping to EXCLUDED is of particular value, as having source data without enough information to map accurately is a tenant of a real world dataset, and provides an additional challenge for machine learning models.

We examined cases where our model’s predictions were incorrect and found that in the majority of cases, the incorrect prediction was extremely similar to the actual label. For example, in one instance the model predicted LOINC 20570-8 which has target term "Hematocrit Bld VFr Pt Qn", and the labeled LOINC was 4544-3 which has target term "Hematocrit Bld VFr Pt Qn Automated count". While these are very similar, because we want to maintain the highest degree of accuracy possible it still counts against the model’s performance. This highlights the importance of having both correct and consistent mappings in our labeled dataset. While we don’t have a measure of the accuracy of the manual mappings that we used to create our labels, in future work we plan to use the incorrect model predictions as a starting point to review the accuracy of existing labeled rows.

Our results are comparable in accuracy to the best reported results from prior studies involving automated laboratory data mapping. The high accuracy and precision of the ensemble learning allows for supplemental use of automated mapping methods alongside clinical terminologists in a range of potential applications. Aside from directly predicting new mappings, this method can also be used as an extra level of quality analysis for maps done manually. Additionally it can be used to verify existing maps, especially those done many years in the past, to identify potentially incorrect mappings to be reexamined. In summary this study has shown that using automated mapping methods has potential to not only reduce manual harmonization time by clinical terminologists, but also to provide an immediate quality improvement to real-world datasets.

The novelties and areas of strength in this study include (a) assessing the place of automated laboratory mapping methods in a real-world dataset, (b) demonstrating the efficacy of such methods on an oncology-specific dataset, (c)
implementation of two novel laboratory free text encoding schemes, and (d) high prediction accuracy using ensemble learning. In future work, we would like to explore other encoding methods, as well as attempt to use multiple encoding methods for different fields within the same model. Furthermore, we would like to explore the extensibility of these techniques to other domains, such as free text medication administration data.

This study has a few notable limitations. Firstly, while the methods in this study are reproducible at no cost, the labeled dataset of oncology laboratory data to LOINC codes used in this study is not publicly available. However, there may be opportunity in the future to open source some of these mappings for the larger research community. A second limitation is that while these encoding and models performed well on oncology data, there may be oncology-specific aspects about laboratory data that would cause these methods to perform poorly in other medical domains. A limitation with our encoding methodology is that the encoded vectors created are very wide, which can lead to long model train times and reduced development speed. Encoding with Levenshtein ratios further increases this issue, as comparing strings is an expensive operation. Another limitation is that because machine learning classifiers can only predict LOINCs that they have been trained on, any new additions to the LOINC catalog or changes in mapping policy would be impossible for the classifier to predict. Whenever this happens all models would need to be retrained on additional applicable training data. Lastly, our primary focus is reducing the burden of clinical terminologists and we are therefore satisfied with high accuracy on a subset of data. Other applications however might require full automation, for which our study has comparably high accuracy, but decreases the applicability of our ensemble learning methods.

Conclusion

As the use of real-world data continues to grow, automated methods that allow for accurate aggregation and harmonization of data from EHRs become increasingly important. Mapping free text laboratory data to LOINC is important before the data can effectively be leveraged for use in research, however the manual mapping process is extremely time intensive. This study has shown that through the use of automated methods, we can significantly lower this burden. Free text medical data is challenging to encode for existing NLP methods, and so we developed specific encoding methodologies to effectively capture and encode the information required to map laboratory data. We demonstrated that with these encodings and the use of ensemble learning, training an automated laboratory data classifier with high accuracy is not only possible, but can provide immediate value to the creation of a high-quality real-world oncology dataset.

References


Human Factors Considerations in Transitions in Care Clinical Decision Support System Implementation Studies.
Erin E. Kennedy, BSN, RN, Kathryn H. Bowles, PhD, RN, FAAN, FACMI
University of Pennsylvania School of Nursing, NewCourtland Center for Transitions and Health Philadelphia, PA

Abstract
Objective: Review transitions in care clinical decision support system (CDSS) implementation studies and describe human factors considerations in users, design, alert types, intervention timing, and implementation outcomes.
Methods: Literature review in PubMed guided by subject matter experts.
Results: Twelve articles were included. Targeted users included physicians, nurses, pharmacists, or interdisciplinary teams. Alerts were deployed via email, cloud-based software, or the EHR in inpatient and/or outpatient settings. Outcome measures varied across articles, with mixed performance. There were six readmissions-focused, two prescribing, one laboratory, two prescribing and laboratory, and one discharge disposition CDSS. Few articles reported statistically significant differences in outcomes, and many reported alert fatigue.
Discussion and Conclusion: Despite the increasing prevalence of CDSS for transitions in care, few articles describe implementation processes and outcomes, and evidence of clinical practice improvement is mixed. Future studies should utilize implementation science frameworks and incorporate appropriate implementation outcomes in addition to traditional clinical outcomes like readmission rates.

Background and Significance
In 2017 alone, 3.5 million potentially preventable 30-day hospital readmissions cost the United States healthcare system $33.7 billion. Readmissions are a multifactorial problem involving patients’ clinical and social history prior to hospitalization, during hospitalization, and in post-discharge life. Readmission reduction remains a central focus in health policy because it holds the potential to improve quality and reduce costs, and discharge planning plays a key role. Discharge planning is complex, individualized, and ultimately determines if the patient will receive any post-acute care (PAC). PAC includes long-term acute care hospitals, inpatient rehabilitation, skilled nursing facilities, and home health care. A human factors study found that patients often have up to six people directly involved in the process, spanning different clinical specialties, family members, and outpatient providers. There are no clinical practice guidelines for this complex process. Variation in clinical decision making can exacerbate health disparities and affect outcomes when patients who need PAC do not receive it.

Coordinated discharge planning and transitions in care from hospital to the post-acute phase are successful readmission reduction strategies, and numerous informatics solutions such as clinical decision support systems (CDSS) have successfully been developed to support this goal. Most early CDSS research prior to 2010 described the benefits of CDSS but did not explore any adverse consequences of this new technology. Early studies found that the use of CDSS improved uptake of consultant’s recommendations by 30%. Bright’s early systematic review of the effects of CDSS systems identified 148 randomized controlled trials and found that the use of commercial CDSS tools improved ordering of laboratory tests, drugs, and preventative services, but did not identify any negative consequences. A study that developed a taxonomy of front-end CDSS tools in commercial and local EHRs found that 53 unique types of CDSS exist, but only 8 were present in all EHR systems. The most common CDSS tools were for medication dosing, and the least common tools were for complicated decisions like discharge planning. As CDSS research became more prolific, articles began to include information about the negative side of CDSS. In 2014, Dhiman et al. attempted to create an evidence-based guide to CDSS purchasing for health systems, but struggled to find much evidence in the literature. The article called for increased transparency from vendors to promote the building of such tools, but many of the barriers that he faced like diversity of EHR systems, legal troubles, limitations in the marketplace, and conflicts of interest continue to exist today.

Although CDSS is ubiquitous in most US health systems, research articles typically focus on variables and model development rather than measuring the effects of CDSS design and implementation on clinicians and patients. Reviews of prediction models that might drive CDSS for discharge planning in acute care including readmissions and discharge disposition have been published, but little information exists about other elements of care transitions like medications, laboratory results, or post-acute settings. The goal of this paper is to review transitions in care CDSS implementation studies and describe human factors characteristics in users, design, alert types, intervention timing, as well as implementation outcomes. Future studies can use this evidence to design more effective implementation strategies.
Methods

Subject matter experts were consulted to refine search terms in PubMed with the following strategy:
(implement* OR satisfaction OR acceptance OR "provider use" OR workflow) AND clinical decision support AND
(discharge planning OR readmission). Articles published before November 1, 2020 were evaluated. Characteristics of
CDSS, outcomes, and insights from the implementation process were extracted and summarized.

Results

The PubMed search returned 373 results. Five additional articles were identified and included for
evaluation from reference lists. Study abstracts were reviewed and 12 were included in the analysis. Inclusion
criteria aimed for studies that implemented CDSS for care transitions from acute care hospitalization to post-acute
settings for adults ≥18 years in the US and described implementation processes. Study selection is described in
Figure 1.

The earliest study was published in 201215 and the most recent were
published in 2020,11, 16, 17
While most studies were
conducted in academic medical centers,15-19 there
were two in regional hospitals,11, 20
and three that included both a hospital and outpatient
facility22-23. Three studies
utilized commercial tools,11, 16,
23 while nine built their
own,15, 17-22, 24, 25 Six studies
tested passive alerts,11, 16, 17, 19,
21, 22 five tested active alerts,15,
20, 23-25 and one18 tested both
types. Study characteristics
are described in Table 1.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Objective</th>
<th>Study Design</th>
<th>Sample</th>
<th>Relevant Clinical Outcomes</th>
<th>Relevant Process Outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bowles et al., 201920</td>
<td>Evaluate effects of CDSS on post-acute care referrals (PAC) and patients’ acute care utilization</td>
<td>Quasi-experimental pre-post design</td>
<td>8,308 hospitalized patients aged 55 years or older from 14 types of units (3,302 control phase, 5,006 intervention phase)</td>
<td>-In control vs. intervention phases, no significant change in PAC referral rates (59.5% vs 59.3%, p=0.55) or referral location (22.8% vs. 23.6% to home health care and 36.6% vs. 35.5% for facility care) -Significant decline in 7 (2.6%, p=0.001), 14 (2.8%, p=0.001), and 30-day readmissions (2.7%, p=0.002) -Characteristics of patients referred to various PAC sites changed between study phases</td>
<td>-CDSS identified 24% and 25.6% more patients for PAC than discharge dispositions in control and intervention phases</td>
</tr>
<tr>
<td>Cochran et al., 201819</td>
<td>Design REDCap Program with risk predictive algorithms (including 30-day readmissions) for pancreaticoduodenectomy</td>
<td>Pretest-posttest design</td>
<td>400 pancreaticoduodenectomy patients at one hospital to design the prediction models; 78 patient (50 pre-implementation, 28 post-implementation) outcomes in the</td>
<td>-Moderate to good discrimination for 30-day readmission (AUC 0.641) and 30-day mortality (AUC 0.856) -30-day readmission rate decreased 15.8% for high-risk patients and 4.4% overall</td>
<td>-Compliance tracking improved from 62.5% to 68.2% after implementation -Adoption: 73.1% of monitored events were action. 67% of actions were record updating. The majority</td>
</tr>
<tr>
<td>Study</td>
<td>Description</td>
<td>Methodology</td>
<td>Outcomes</td>
<td>Implications</td>
<td></td>
</tr>
<tr>
<td>-----------------------</td>
<td>--------------------------------------------------------------------------------------------------</td>
<td>-------------------------------------------------------------------------------</td>
<td>---------------------------------------------------------------------------</td>
<td>------------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>Gallagher et al., 2020</td>
<td>Describe implementation of Epic system 30-day readmission risk model in acute care</td>
<td>Observational study</td>
<td>-Mean length of stay decreased from 12.8 to 10.1 days (p=0.033)</td>
<td>-Projected postoperative laboratory test costs decreased approximately $150,000 in 5 months post-implementation of users (62.5%) completed actions in ≤ 43 seconds</td>
<td></td>
</tr>
<tr>
<td>Hewner et al., 2017</td>
<td>Introduce social determinants of health (SDH) into web-based patient centered assessment method (PCAM) to improve care coordination for chronically ill individuals during care transitions from hospital to home</td>
<td>Pretest-posttest design</td>
<td>-Significant difference in inpatient utilization rates (chi-squared=5.07, p=0.02), outpatient utilization (W statistic=48.661, p=0.0025), and ED utilization (W statistic=64172, p&lt;0.001) in pretest vs. posttest periods</td>
<td>-Fidelity: to onboard, care coordinators needed assistance understanding the importance of assessing behavioral health issues and developing a plan of care to address problems -Main benefit in alerts is in training the staff in population-based care -Initially planned to implement in EHR but decided to use ONC standards for interoperability in the regional health information exchange -Staff took several months to see value in intervention</td>
<td></td>
</tr>
<tr>
<td>Hewner et al., 2018</td>
<td>Demonstrate feasibility of Coordinating Transitions Intervention (CTI) CDSS in primary care to reduce hospitalizations during care transitions from hospital to home</td>
<td>Pretest-Posttest design</td>
<td>-Inpatient hospitalizations decreased by 25% (p=0.09), and ED utilization decreased by 35% (p=0.001) in the post-test period -Outpatient visits increased by 27% (p&lt;0.001) -Reduction in inpatient and ED encounters led to $1,669 less per Medicaid recipient, with $664 related to the intervention -Outpatient visits generated $71,289 in new revenue</td>
<td>Described in Hewner et al., 2017</td>
<td></td>
</tr>
<tr>
<td>Horne et al., 2020</td>
<td>Evaluate the impact of a health system readmission and mortality risk reduction program with CDSS and risk-score guided multidisciplinary team-based care process (MTCP) among heart failure patients</td>
<td>Phased implementation with sequential crossover in a stepped manner</td>
<td>-Compared with controls, MTCP recipients had 21% lower 30-day readmission (p=0.13) and 52% lower 30-day mortality (p&lt;0.001) -Low-risk patients did not experience increased readmissions or mortality -No significant change in length of stay or costs by MTCP</td>
<td>-Implemented interdisciplinary process for inpatient care, discharge planning, and post-discharge follow-up in the high-risk patient group -MTCP patients had significantly higher proportion of home health discharge disposition compared with controls (27.8% vs 22.6%, p=0.07)</td>
<td></td>
</tr>
<tr>
<td>Romero-Brunau et al., 2020</td>
<td>Reduce unplanned readmissions through the use of artificial</td>
<td>Pretest-posttest design</td>
<td>-Model sensitivity and specificity were 65% and 89% - Statistically significant absolute (3.3%, p&lt;0.001) and relative (25%, p&lt;0.001)</td>
<td>-Most useful component was ability to identify and prioritize high-risk patients</td>
<td></td>
</tr>
<tr>
<td>Medication/ Laboratory CDSS</td>
<td>Amroze et al., 2019&lt;sup&gt;22&lt;/sup&gt;</td>
<td>Evaluate feasibility of using EHR audit logs to track opening of and response to non-interruptive medication and laboratory alerts in the EHR in primary care among primary care providers</td>
<td>Descriptive study following a clinical trial (Gurwitz et al., 2014&lt;sup&gt;21&lt;/sup&gt;)</td>
<td>799 non-interruptive alerts from 75 primary care providers</td>
<td>-See Gurwitz et al., 2014&lt;sup&gt;21&lt;/sup&gt; -Alert types: 74.2% information only, 4.6% medication recommendations, 21.1% laboratory test recommendations -78.5% of alerts opened by the correct physician and 33% prompted immediate action -After including other providers, additional 41.3% of alerts had follow-up actions by the end of the next day and 32.7% had none -Odds of immediate action after a recommendation alert were higher vs. information only</td>
</tr>
<tr>
<td>Blecker et al., 2019&lt;sup&gt;14&lt;/sup&gt;</td>
<td>Compare effectiveness and implementation of interruptive vs. non-interruptive CDS to improve angiotensin converting enzyme (ACE) inhibitor continuation at discharge</td>
<td>Pseudo-randomized design (patient-level)</td>
<td>958 hospitalized heart failure patients (465 interruptive group, 493 non-interruptive group)</td>
<td>-Interruptive alerts had higher discharge utilization rates than non-interruptive (79.6% vs. 74.2%, p=0.05) -Overall ACE inhibitor or ARB utilization rates for eligible patients increased after implementation (79.6% vs 73.6%, p=0.04) -Adoption: Interruptive alerts generated more responses (40.6% vs 13.1%, p&lt;0.001), more contraindications reported (33.1% vs. 11.3%, p&lt;0.001) -Fidelity: Interruptive alerts generated more ACE inhibitors ordered in 12 hours (17.6% vs 10.3%, p&lt;0.01) -Interruptive alert triggered median of 14 times per hospitalization with response rate 1.7%</td>
<td></td>
</tr>
<tr>
<td>Dalal et al., 2012&lt;sup&gt;15&lt;/sup&gt;</td>
<td>Describe the design and implementation of automated email notification of tests pending at discharge (TPAD) to inpatient-attending physician and facilitate communication with primary care physician (PCP)</td>
<td>Pilot study</td>
<td>178 hospitalized cardiology and general medicine patients (83 patients in phase I, 95 patients in phase II) and 29 physicians in user satisfaction surveys</td>
<td>-Not studied -System correctly identified 98.8% of patients and 96.4% of physicians -18% of emails flagged abnormal results -Inpatient physicians received 1.6 notifications per patient with TPAD (range 1-31 emails) -PCP was only copied on email in 52% of cases -84% of inpatient physicians were satisfied or very satisfied with intervention</td>
<td></td>
</tr>
<tr>
<td>Elliot et al., 2017&lt;sup&gt;23&lt;/sup&gt;</td>
<td>Assess clinical impact of pharmacogenetic profiling integrating binary and cumulative drug and gene interaction warnings on home health polypharmacy patients</td>
<td>Randomized controlled trial</td>
<td>110 hospitalized patients aged 50 years or older with polypharmacy and home health care discharge dispositions (57 intervention, 53 control)</td>
<td>-Non-significant reductions in 30-day readmissions (mean hospitalizations per patient in intervention vs. control group 0.25 vs. 0.38, p=0.21) or ED visits (0.25 vs. 0.40, p=0.16) -Significant reductions in 60-day readmissions (intervention vs control group 0.33 vs. 0.70, p=0.007) and ED utilization (0.39 vs. 0.66, p=0.045) -Time to first rehospitalization hazard ratio (HR) was 0.59 and time to first ED visit was 0.60 -In intervention group, mean number of YouScript vs. gene-based recommendations was 2.18 vs. 1.49 -Of the recommendations passed to clinicians, 77% were followed, 5% were not followed, and 18% were unknown</td>
<td></td>
</tr>
</tbody>
</table>
The CDSS systems targeted discharge disposition, readmission, medications, and laboratory tests. Implementation outcomes varied across articles. Table 2 describes the human factors and implementation characteristics discussed in the review. Each result will be discussed in two groups: readmission and discharge disposition prediction tools, as well as medication and laboratory tools.

<table>
<thead>
<tr>
<th>Intervention Timing</th>
<th>Human Factors and Implementation Characteristics for Transitions in Care-Focused CDSS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inpatient Alerts</td>
<td>- Throughput entire hospitalization[^11^, ^16^, ^17^, ^19^]</td>
</tr>
<tr>
<td></td>
<td>- Twice-daily from 48-hours after admission until discharge[^20^]</td>
</tr>
<tr>
<td>Transitional Care Alerts</td>
<td>- Immediately after discharge[^24^, ^25^]</td>
</tr>
<tr>
<td>User</td>
<td>- Nurse[^24^, ^25^]</td>
</tr>
<tr>
<td></td>
<td>- Interdisciplinary[^11^, ^16^, ^17^, ^20^]</td>
</tr>
<tr>
<td></td>
<td>- Not specified[^29^]</td>
</tr>
<tr>
<td>Alert Type</td>
<td>- Passive EHR alert[^16^, ^17^]</td>
</tr>
<tr>
<td></td>
<td>- Passive web-based alert[^11^, ^19^]</td>
</tr>
<tr>
<td></td>
<td>- Email alert[^24^, ^25^]</td>
</tr>
<tr>
<td>Design</td>
<td>- Color coded button next to patient name with expandable screen</td>
</tr>
<tr>
<td></td>
<td>- Containing additional information[^1^, ^16^, ^17^]</td>
</tr>
<tr>
<td></td>
<td>- Email containing risk scoring and recommendation[^20^, ^24^, ^25^]</td>
</tr>
<tr>
<td></td>
<td>- REDCap dashboard displaying patient risk category and risk terciles for each outcome[^19^]</td>
</tr>
<tr>
<td>Outcomes Studied</td>
<td>- Adoption[^25^]</td>
</tr>
<tr>
<td></td>
<td>- Model performance[^11^, ^16^, ^19^]</td>
</tr>
<tr>
<td></td>
<td>- Readmissions[^10^, ^16^, ^17^, ^19^, ^20^, ^24^, ^25^]</td>
</tr>
<tr>
<td></td>
<td>- Emergency department utilization[^20^, ^24^, ^25^]</td>
</tr>
<tr>
<td></td>
<td>- Length of Stay[^17^, ^19^]</td>
</tr>
<tr>
<td></td>
<td>- Mortality[^17^, ^19^]</td>
</tr>
<tr>
<td></td>
<td>- Outpatient visits[^24^, ^25^]</td>
</tr>
<tr>
<td></td>
<td>- Cost[^1^, ^18^, ^24^, ^25^]</td>
</tr>
<tr>
<td></td>
<td>- Ratio of risk score to interventions performed[^16^]</td>
</tr>
<tr>
<td></td>
<td>- Qualitative Feedback[^11^, ^24^, ^25^]</td>
</tr>
</tbody>
</table>

Legend: Columns represent the two main categories of CDSS alerts studied in this review (readmission-risk and discharge disposition CDSS vs. medication and laboratory CDSS). Rows represent categories of human factors consideration, which are discussed in more detail below.

**CDSS Description, Intervention Timing, and Alert Type**

Care transitions are multi-faceted and interdisciplinary, so the CDSS in this review target different aspects and actors in the process. Six articles implemented tools to predict 30-day hospital readmissions[^11^, ^16^, ^17^, ^19^, ^24^, ^25^] two implemented prescribing tools[^18^, ^23^] one implemented a tool to predict discharge disposition[^20^] one implemented laboratory alerts[^15^] and two articles from the same study implemented both prescribing tools and laboratory tools[^21^, ^22^]. These tools were designed to improve discharge planning, communication between inpatient and outpatient providers, post-acute care, or a combination of settings. Although most CDSS were only implemented in one setting, some leveraged health system EHR integration or regional health information exchanges.

Most readmission prediction tools were designed to be used during the hospitalization. Gallagher et al.[^16^] argued that ideal readmissions prediction tools are “closely integrated with the electronic health record; pull data real-time from the EHR, update continuously, and present the score for clinical teams to act upon.” Their tool was built as a passive alert available in the EHR from admission to discharge for all clinicians with a focus on physicians and case managers. The team conducted quarterly interdisciplinary meetings over two years to monitor the implementation[^16^]. The Horne et al.[^17^] tool was designed similarly, with a daily EHR report for interdisciplinary rounds within 24-hours of admission to discharge with specific tasks for bedside nurses, charge nurses, pharmacists,
social workers, dieticians, care coordinators, and heart failure clinic nurse providers in a multidisciplinary team-based care process (MTCP) in the inpatient setting only. The Romero-Brufau et al.\textsuperscript{11} tool was not implemented in the EHR due to information technology resource limitations, but it was available in an external web portal to discharge planners and physicians from admission to discharge as a passive alert, similar to Gallagher et al. Their team conducted workflow mapping for six months prior to implementation and conducted informal qualitative interviews with physicians in the post-implementation period. Similarly, Cochran et al.\textsuperscript{19} developed and implemented their readmission risk-prediction tool in a web-based platform. The dashboard was used throughout the inpatient hospitalization, but authors did not specify the frequency or types of clinician users. They also developed prediction models for mortality and serious complication.\textsuperscript{19} Bowles et al.\textsuperscript{11} implemented a CDSS tool to predict discharge disposition, and tested its impact on readmission. Their CDSS was an active twice-daily alert in the form of an email to discharge planning clinicians (case management). The study team worked extensively with case management to design the intervention prior to the intervention phase.

Hewner et al.\textsuperscript{24, 25} took a different approach to readmission prediction by implementing the Patient-Centered Assessment Method (PCAM) platform across the transition from hospital to the outpatient setting, and made the tool interoperable throughout other health systems in the region with an EHR-agnostic web-based platform. The clinical data repository received hospitalization information, computed a readmission risk profile based on comorbidities and Medicaid status which triggered a CDSS alert to the outpatient nurse case manager to call the high-risk patients within 48 hours of discharge, complete the PCAM questionnaire, and intervene as necessary. This was the only article targeting nurses only. The team designed the intervention over 6 months prior to implementation, assessed implementation for 12 months, and continued to monitor the intervention after study completion.\textsuperscript{24, 25}

Two articles reported on CDSS tools focused on medications, and one focused on the hospitalization period. Blecker et al.\textsuperscript{18} developed and implemented a CDSS tool to increase prescriptions of ACE inhibitors at discharge among heart failure patients and tested both passive and active alerts in the pre-post study. The alert was designed for use by physicians only and was available in the Epic EHR system throughout the entire hospitalization even though the focus of the study was the discharge medication list. Authors claim to have used user-centered design in development and testing but did not provide details or collect qualitative data. In addition to heart failure diagnosis, CDSS logic incorporated ejection fraction, medication list, blood pressure, and pregnancy status.\textsuperscript{18}

Similar to the Hewner et al.\textsuperscript{24, 25} articles for readmissions, Elliot et al.\textsuperscript{23} designed a pilot laboratory intervention to recruit patients at discharge and deploy the YouScript CDSS intervention targeting older polypharmacy patients with potential drug-gene based interactions during home care. Their team did not use user-centered design or collect any qualitative data, but the intervention was designed for pharmacists only and included automated prescription recommendations for patients in the intervention group. Dalal et al.\textsuperscript{15} designed and implemented a laboratory CDSS tool to notify both the inpatient physician and outpatient primary care provider (if in the same health system and present in the EHR) of pending laboratory results at discharge. The study was designed to address the chasm between inpatient and outpatient settings, because lack of follow up to tests pending at discharge can contribute to readmissions, delays in diagnosis or treatment, and adverse patient outcomes. Even though this was one of the earliest articles in the review, its design truly bridged the gap between inpatient and outpatient settings. The team applied user-centered design in the pre-implementation period and the intervention was deployed to inpatient physicians and outpatient providers after discharge to facilitate communication and patient handoff.\textsuperscript{15} Gurwitz et al., 2014 assessed the impact of primary care EHR-based passive alerts (development described elsewhere)\textsuperscript{26} about drug-drug interactions, medication recommendations, and laboratory recommendations for older adult patients 3 days following discharge to the community in a randomized controlled trial. Amroze et al., 2019\textsuperscript{22} later analyzed the alert types, opening of, and response to these alerts. Although the original trial did not use user-centered design, Amroze et al.\textsuperscript{22} did perform simulation testing with a physician and data manager.

**CDSS Design**

All of the CDSS tools in this review were implemented electronically, either embedded in EHRs, external portals, or emails. EHR and email were the most common alert types across articles. Although most were implemented in inpatient hospital settings only, others were designed to facilitate communication with post-acute care or outpatient settings. For readmissions on the inpatient side, the Gallagher et al.\textsuperscript{16} readmission prediction tool was implemented as a column called “Readmission Risk” in Epic work lists with a red (high risk), yellow (medium risk), or green (low risk) circle that the clinician can click to see an expanded view of their readmission risk percentage, factors contributing to score in order of importance, and changes to readmission risk over time. The score was updated every four hours and the tool was not customized to their health system. Similarly, the Horne et al.\textsuperscript{17} readmission-risk daily report (passive alert) was run at 9:00am each day in the EHR which classified patients...
into high- or low-readmission or mortality risk. Nurses later reviewed the lists and emailed them to pharmacists and charge nurses each day to begin the MTCP pathway, with a separate process for weekends and holidays. Although it was not possible for the Romero-Brufau et al.\textsuperscript{11} team to implement their tool in the EHR, their portal was designed similar to the Gallagher et al. team with a purple dot next to patients’ names to identify those in the top 20\textsuperscript{th} percentile of readmission risk that clinicians could click to view the risk factors contributing to the score and up to 26 possible recommendations for interventions. The score was updated every 24 hours each morning before shift change, and discharge planners were instructed to contact physicians daily to discuss the recommendations. Cochran et al.\textsuperscript{19} developed an interface in REDCap with five data collection forms related to patient characteristics and Enhanced Recovery After Surgery protocol reporting requirements, and output forms with probability, clinical, and cost predictions for clinicians to improve decision making. Risk categories included high, low, and expected.\textsuperscript{19} Other discharge disposition and readmission-risk CDSS tools were implemented via email. Hewner et al.\textsuperscript{24, 25} implemented the CDSS alert for patients at high-risk of readmission as a secure email to the outpatient nurse case manager with information about the patient, hospitalization, and pre-existing comorbid conditions. Initially the PCAM assessment was completed on paper, but clinicians were so satisfied with the tool that they digitized it by the end of the study. Bowles et al.\textsuperscript{20} fired the CDSS algorithm recommendation for discharge disposition as a twice-daily report via email to the inpatient discharge planning team along with characteristics associated with the recommendation.

The medication and laboratory CDSS followed similar approaches with EHR-integration and emails in acute care and/or post-acute settings. The Blecker et al.\textsuperscript{18} team aimed to increase discharge prescriptions of ACE inhibitors among HF patients and implemented CDSS within Epic into existing prescribing workflows. Rather than delivering the alert at discharge, the team tested two versions of the CDSS throughout the entire hospitalization to see if it would lead to the continuation of the ACE inhibitor on the discharge medication orders. The active alert intervention group received a pop-up in the ordering screen that stated the patient was not on an ACE inhibitor, recommended the therapy, and included the patient’s blood pressure, creatinine, potassium, eGFR, and ejection fraction. Providers had the option to mute the alert every six hours. The passive alert was located in the daily provider checklist with the same explanation and details as the active alert, but never popped up or interfered with the clinician’s EHR workflow.\textsuperscript{18} Elliot et al.\textsuperscript{21} reported delivering the YouScript CDSS to home health care pharmacists to reduce drug-gene based interactions among older polypharmacy patients. However, they did not provide any details or screenshots of their CDSS intervention, nor did they state the frequency of reports. They claimed that the reports contained drug-gene interactions and recommendations with no further details. The Dalal et al.\textsuperscript{15} laboratory CDSS tool was implemented as one email per test grouping (chemistry, biology, etc.) per patient per day to each inpatient and outpatient provider. They designed customizable settings so that physicians could suppress certain notifications and added a feedback link within the email to address provider concerns in real time. The medication and laboratory CDSS from Gurwitz et al., 2014\textsuperscript{21} and Amroze et al., 2019\textsuperscript{22} were implemented as noninterruptive EHR InBasket alerts to primary care providers of patients aged 65 and older 3 days after discharge from acute care to the community. The alert types were classified as information only (drug-drug interaction), medication recommendation, or laboratory recommendation, while the control group received usual care.\textsuperscript{21, 22}

\textbf{Outcomes Studied}

Most articles in the review reported clinical outcomes rather than implementation outcomes. The gold standard in implementation science research is to select implementation outcomes based on frameworks such as RE-AIM\textsuperscript{27} or PRISM.\textsuperscript{28} Although all of the articles completed implementation, only two articles in this review cited a specific implementation framework (Proctor’s taxonomy\textsuperscript{19} and the Quadruple Aim\textsuperscript{28}). This aligns with prior research demonstrating a lack of theoretical underpinnings in implementation science research as a whole.\textsuperscript{29} Therefore, the implementation-focused outcomes will be described as process outcomes in this review and Table 1.

For the readmissions and discharge disposition-focused CDSS, clinical outcome measures included model performance, readmissions, mortality, emergency department utilization, cost, and outpatient visits. Although model performance was fair to good in these articles, the implementation outcomes demonstrated mixed success. Gallagher et al.\textsuperscript{16} reported a stable area under the curve (AUC) of 0.716–0.760 for their readmission risk prediction tool in overall adult patients from pre-implementation to two-years post-implementation, but no statistically significant change in 30-day readmissions or number of interventions performed relative to risk score. The Romero-Brufau et al.\textsuperscript{11} tool which predicted readmission risk and recommended interventions reported model sensitivity and specificity of 65% and 89%, respectively and a 25% relative (3% overall) reduction in readmission rates over the six-month implementation period. Authors argued for setting a low sensitivity threshold to reduce alert fatigue. For process outcomes in the Romero-Brufau et al.\textsuperscript{11} article, staff expressed frustration that readmission reduction interventions in the CDSS were inappropriate and the tool would be better off without the clinical recommendations. Authors noted that this is consistent with other studies that showed artificial intelligence tools perform well with
discrete tasks (like readmission risk) but poorly with complex tasks like care planning.\textsuperscript{11} Horne et al.\textsuperscript{17} reported 21\% lower readmission and 52\% lower 30-day mortality (both statistically significant) among high-risk MTCP patients compared with high-risk controls, and no significant increases in either outcome among lower-risk patients. The team successfully implemented the MTCP across health systems and identified a process outcome of significantly higher proportion of home healthcare discharge disposition compared with controls (24\% vs 22.6\%).\textsuperscript{17} For clinical outcomes, Cochran et al.\textsuperscript{19} reported moderate to good model discrimination for 30-day readmissions (AUC 0.641) and 30-day mortality (AUC 0.856), readmissions decreased 15.8\% among high-risk patients after implementation (4.4\% overall), mean length of stay decreased by 2.7 days, and post-operative laboratory test costs were projected to decrease $150,000 in 5 months post-implementation. For process outcomes, compliance tracking improved by 5.7\%. In terms of adoption, 73.1\% of monitored events in the application were browsing and 26.9\% were action, of which 67\% were record updating, and the majority of users completed actions in ≤43 seconds.\textsuperscript{19} The Hewner et al.\textsuperscript{24,25} readmission prediction CDSS for outpatient case managers measured changes in emergency department utilization, 30-day readmissions, outpatient utilization, and cost. Their transitional care intervention reduced emergency department visits by 35\%, 30-day readmissions by 25\%, and increased outpatient visits by 27\%, which led to an overall savings of $664 per patient and $71,289 in overall revenue from outpatient visits.\textsuperscript{24} Although authors did not conduct a rigorous qualitative study, they reported improvements in patient experience, work life of clinicians, and value of the nurse. The Bowles et al.\textsuperscript{20} discharge disposition CDSS evaluated the impact on readmissions and found a statistically significant 22\% relative reduction (3\% overall) in 30-day readmissions, but no statistically significant impact on 30-day emergency department utilization. For process outcomes, they found that the CDSS recommended 24-25.6\% more patients for post-acute care than clinicians.\textsuperscript{20}

For the medication-focused CDSS, clinical outcome measures included readmissions and process outcomes included adoption and fidelity. Blecker et al.\textsuperscript{18} defined adoption as the “percentage of hospitalizations with at least one alert that included an action to an alert that was not a dismissal” and fidelity as “the percent of hospitalizations with at least one alert that led to signing an order.” The active alerts triggering ACE inhibitor prescription for HF patients in acute care had higher adoption (40.6\% alerts were addressed) compared to passive alerts (13.1\% were addressed), but there were no statistically significant differences in fidelity between intervention groups. The alerts were triggered nearly 26 times per hospitalization (response rate <2\%), the 12-hour alert ordering rate was 0.8\%, and the alert was fired 405 times for each ACE inhibitor to be prescribed. Although alert fatigue was not discussed, it is likely that this CDSS tool was contributing to physician alert fatigue.\textsuperscript{18} The tool did not have any feedback channel for end-users, so a future study might consider adding capability for feedback, altering alert frequency, and/or conducting qualitative interviews in the post-implementation period. The Elliot et al.\textsuperscript{23} home care polypharmacy CDSS intervention did not lead to significant reductions in 30-day readmissions, but there were significant differences in 60-day readmissions (0.33 vs 0.70 per patient in tested vs untested group). Differences in OASIS quality outcomes including overall health status, pain, confusion, anxiety, depression, disruptive behavior, and assistance with activities of daily living were non-significant between groups. Authors reported 77\% of drug therapy recommendations that went to clinicians were followed.\textsuperscript{23}

Outcomes in the pending laboratory results CDSS tool included physician satisfaction and number of alerts. In the Dalal et al.\textsuperscript{15} study of emailing pending laboratory results at discharge to inpatient and outpatient providers, they reported 1.6 emails per patient and 84\% satisfaction among the physician end users. They credited physician feedback for their ability to reduce possible alert fatigue and achieve high satisfaction rates. For the medication and laboratory-focused CDSS in primary care after hospital discharge, clinical outcomes included outpatient visit rates within 30-days of discharge (27.7\% for intervention and 28.3\% for the control group) and 30-day readmissions (18.8\% for intervention and 19.9\% in the control group), both of which were not significantly different.\textsuperscript{21} This unexpected result prompted the follow-up article about types, opening of, and response to alerts, which are process outcomes. Authors reported 74.2\% of alerts were information only, 4.6\% were medication recommendations, and 21.2\% were laboratory recommendations. 78.5\% of alerts were opened by the target physician and 33\% prompted immediate action. After including other providers in the practice and expanding the time point to the end of the following day, an additional 41.3\% of all alerts had follow-up action and 32.7\% had no action. Odds ratios for action by the end of the following day were higher for medication and laboratory recommendations compared with information only. Authors identified two implementation challenges (capturing and interpreting historical data elements) and recommended that future studies plan how to collect EHR audit data prior to implementing an intervention.\textsuperscript{22}

**Limitations**

The search was limited to articles referenced in PubMed. Limited search terms may have missed articles. Future studies should include additional databases and broader search terms. The articles were not evaluated for quality.
Conclusion

Over the last decade, healthcare policymakers and administrators have invested in more predictive analytics tools to target various aspects of discharge planning to reduce hospital readmissions and costs. While many of these tools have been implemented as CDSS in health systems, the published academic literature on this topic is sparse. This review represents a starting point summarizing the current state of implementations to assist potential developers and purchasers of CDSS in their design and implementation processes.

The CDSS tools in this review were designed to reduce readmissions and improve prescribing or laboratory follow up with great variation in designs and outcomes studied. End-users varied with the majority targeted at physicians only, while others focused on nurses, pharmacists, or interdisciplinary discharge teams. While most tools were implemented in hospitals, some incorporated inpatient and/or outpatient settings. Design characteristics included alert frequency, EHR integration, information included in the alert, and active versus passive alerts. Email was a common alert type, which might be the result of workarounds in health systems where EHR-integrated CDSS capability is limited. A CDSS taxonomy study in 2011 called for increased interoperability and customized alerts for different clinical situations. Unfortunately, this remains a barrier in 2020 and only two articles in this review addressed interoperability with mixed success and one article focused on customizable alerts. All of the articles in this review tempered their positive results with limitations and described areas for future improvement. Very few articles reported statistically significant differences in outcomes after the post-implementation period and alert fatigue was a major focus in the discussion sections. Only two articles mentioned outcomes based on implementation science frameworks (adoption and fidelity), and most articles measured traditional clinical outcomes like readmissions. Future studies should focus on the implementation process with special consideration for the CDSS users, design, alert fatigue, and incorporate implementation science outcomes in addition to traditional clinical outcomes.

References


26. Field TS, Garber L, Gagne SJ, et al. Technological resources and personnel costs required to implement an automated alert system for ambulatory physicians when patients are discharged from hospitals to home. Inform Prim Care. 2012;20(2):87-93. DOI: 10.14236/jhi.v20i2.29


Intrinsic Evaluation of Contextual and Non-contextual Word Embeddings using Radiology Reports

Mirza S. Khan, MD¹²³, Bennett A. Landman, PhD²³, Stephen A. Deppen, PhD³, Michael E. Matheny, MD, MS, MPH¹³
¹US Dept. of Veterans Affairs, Nashville, TN; ²Vanderbilt University, Nashville, TN; ³Vanderbilt University Medical Center, Nashville, TN

Abstract

Many clinical natural language processing methods rely on non-contextual word embedding (NCWE) or contextual word embedding (CWE) models. Yet, few, if any, intrinsic evaluation benchmarks exist comparing embedding representations against clinician judgment. We developed intrinsic evaluation tasks for embedding models using a corpus of radiology reports: term pair similarity for NCWEs and cloze task accuracy for CWEs. Using surveys, we quantified the agreement between clinician judgment and embedding model representations. We compare embedding models trained on a custom radiology report corpus (RRC), a general corpus, and PubMed and MIMIC-III corpora (P&MC). Cloze task accuracy was equivalent for RRC and P&MC models. For term pair similarity, P&MC-trained NCWEs outperformed all other NCWE models ($\rho$ spearman 0.61 vs. 0.27-0.44). Among models trained on RRC, fastText models often outperformed other NCWE models and spherical embeddings provided overly optimistic representations of term pair similarity.

Introduction

Quantitative evaluation of natural language processing (NLP) models can generally be categorized into intrinsic and extrinsic evaluation methods. Intrinsic evaluation reflects the correlation between the algorithms and human judgment. This may include testing for syntactic or semantic relationships between words. While much emphasis in NLP-related research is on extrinsic evaluation of NLP methods, it is vital to conduct rigorous intrinsic evaluation. For instance, intrinsic evaluation using word vector analogies has highlighted gender, racial and religion-based biases in word embeddings trained using Google News or Reddit corpora.¹²

Often in the context of NLP methods, intrinsic evaluation tasks judge the similarity or relatedness of pre-selected word pairs. Semantic relatedness or semantic similarity based upon human judgment is often computed as a correlation coefficient by comparing aggregate human assessed scores to similarity metrics, such as cosine similarity, using word vectors. The models in greater agreement to human judgment are considered better models.³–⁵ By contrast, extrinsic evaluation relates to the effect of word embedding input features on the performance of downstream tasks, such as document classification.⁶

Existing clinical benchmarks comparing similarity or relatedness of clinical concepts include Hliaoutakis’ work, MayoSRS, original and modified UMNSRS.⁶–¹⁰ The concept pairs that comprise each of these benchmarks were manually curated by a physician from existing medical ontologies, such as Medical Subject Headings (MeSH) or Unified Medical Language System (UMLS). To assess human judgment of concept pair similarity or relatedness, the researchers conducted surveys of medical coders or clinicians. These benchmarks are publicly available and have been used by others to study the effects of different training corpora or other training strategies.¹⁰–¹³ Importantly, these were created before word2vec and subsequent non-contextual word embedding (NCWEs) models were introduced.

In the general NLP domain, additional intrinsic evaluation tasks have been proposed. The use of pre-specified analogies is a popular approach for intrinsic evaluation of embedding models.¹⁴–¹⁵ Schnabel et al. proposed the comparative intrinsic evaluation task and also recommend other methods for intrinsic evaluation of word embedding models: relatedness, analogy, categorization and selectional preference.¹⁵ Ye and Fabbri use a similar design to the comparative intrinsic evaluation task for clinical NLP assessment.¹⁶

Contextual word embedding (CWE) models, such as BERT, use a different training objective than NCWEs. Specifically, BERT-based models include a masked-language modeling (MLM) pre-training strategy.¹⁵ Given the different training objective of CWEs, Devlin suggests that these models may not provide meaningful cosine similarity measurements.¹⁶ Researchers have proposed intrinsic evaluation tasks in keeping with the BERT modeling approach.
Goldberg and Ettinger respectively evaluate BERT models using previously published cloze (‘fill-in-the-blank’) tasks and compare the ‘gold-standard’ masked word with BERT word predictions.

Few available clinical NLP benchmarks for CWEs exist. The Biomedical Language Understanding Evaluation (BLUE) benchmark contains 5 extrinsic evaluation tasks: sentence similarity, named entity recognition, relation extraction, document classification and inference. We were unable to find any publicly available intrinsic evaluation tasks designed to examine CWEs for the clinical or biomedical domains.

Here, we present our work on the development of intrinsic evaluation tasks for NCWEs and CWEs from a radiology report corpus. We then use these tasks to study the agreement between clinician judgment and embedding models trained using different algorithms or training corpora.

**Methods**

**Study Cohort/Corpus:** We curated a corpus of radiology reports from the Vanderbilt University Medical Center (VUMC) Research Derivative, an extract of the electronic health record from legacy and Epic record data, normalized to the OMOP common data model. We selected only those radiology reports with study descriptions corresponding to computed tomography (CT) scans inclusive of the chest. Report collection was limited to those for patients 18 years of age or older. This study was approved by the VUMC Institutional Review Board.

Text from the publicly available Fleischner Glossary of Terms for Thoracic Imaging was also collected to represent a corpus containing terms more relevant within radiology reports.

![Algorithm for data-driven generation of term pairs from a radiology report corpus using custom non-contextual embedding models.](image)

**Figure 1:** Algorithm for data-driven generation of term pairs from a radiology report corpus using custom non-contextual embedding models.
**Pre-processing:** Pre-processing of the raw clinical text included removal of unicode characters and HTML parsing to plain text using the Beautiful Soup library. We concatenated hyphenated terms by converting '-' characters to '\-', e.g. ‘status-post’ was converted to ‘status_post.’ We also scrubbed dates and times, age, medical record numbers, email addresses, phone numbers, social security numbers and location information. Methods for de-identification included custom regular expressions, the spaCy library tokenizer and our own customizations to the presidio Python package. We also used spaCy for text tokenization using the provided blank English model. Stop words for our radiology report corpus were not excluded. We pre-processed text from the Fleischner glossary using the same process above except that we also performed stop word exclusion using the spaCy English stop word vocabulary.

**Custom NCWEs:** We used the gensim library implementations of the word2vec and fastText models. Spherical embedding models were created using the source code provided by the authors. Each of our word2vec and fastText models were trained using the skip-gram method. We used similar training parameters for each of our custom embedding models: fixed dimension of 200, initial learning rate of 0.025, context window of 5, excluded words occurring fewer than 5 times, sampling threshold of 0.001, negative sampling rate of 5 and trained for 10 epochs.

**Public NCWEs:** We used BioWordVec, which is a 200-dimensional fastText model trained using both a PubMed corpus and text from the Medical Information Mart for Intensive Care (MIMIC-III) dataset and made available by the authors. Additionally, we used a publicly available gensim word2vec model trained using the skip-gram method on the English Wikipedia corpus (February 2017) obtained from the University of Oslo’s Nordic Language Processing Laboratory word vector repository. We also used a spherical embedding model trained on the English Wikipedia corpus provided by Meng et al.

**Custom CWEs:** The pre-processed text of our radiology report corpus was exported to a txt file with each document per line. We used the transformers library to create a custom tokenizer and for pre-training on our radiology report corpus. For training our custom model, we used the same configuration parameters as the original BERT model. Pre-training our custom model took approximately 2.5 days using 2 12GB GPUs.

**Public CWEs:** We compare our custom BERT-based model to the original BERT model and a publicly available BERT-based model pre-trained using a PubMed and MIMIC-III corpus, BlueBERT. Each of these models was downloaded from the HuggingFace Models repository.

**Term pair selection:** We compared the pre-processed Fleischner Society Glossary tokens with the vocabulary of each of our 6 NCWE models. We retained only those terms shared between the Fleischner Society Glossary that were present in all 6 embedding vocabularies. The resulting 1,059 shared terms (‘query words’) were manually reviewed by a physician (MSK), and a subset of these terms (326 terms) were manually selected for possible inclusion in our survey. For each of the 326 query words, we randomly select the $k$th most similar term using each of our 3 custom embeddings, where $k \in \{1, 5, 50\}$. This method was adapted from Schnabel et al. As the examples in Table 1 show, for each query word, we retrieved 3 candidate terms. One of these three was then selected to construct a term pair. A physician (MSK) then manually reviewed the three candidate terms, i.e. the $k$th most similar term to the query word for each of our custom embedding models. For each of our 326 query words, if 2 or 3 of the embedding generated candidate terms were similar, we selected this term to form a term pair. If all 3 of the proposed candidate terms was different, a candidate term was arbitrarily selected by MSK. As we used our custom embeddings to generate term pairs, some of the selected terms for each query word were not present in the publicly available embeddings. For accurate comparison and generation of correlation coefficients, we excluded those term pairs where the word was not present in each of the 6 embeddings being evaluated.

**Table 1:** Example use of the $k$th most similar term from custom embeddings to inform term pair creation.

<table>
<thead>
<tr>
<th>Query Word</th>
<th>k</th>
<th>word2vec</th>
<th>fastText</th>
<th>spherical embedding</th>
<th>Manual Review Selection</th>
</tr>
</thead>
<tbody>
<tr>
<td>abscess</td>
<td>1</td>
<td>empyema</td>
<td>empyema</td>
<td>osteomyelitis</td>
<td>empyema</td>
</tr>
<tr>
<td>adenocarcinoma</td>
<td>50</td>
<td>peritonei</td>
<td>bronchoalveolar</td>
<td>status_post</td>
<td>bronchoalveolar</td>
</tr>
</tbody>
</table>

**Cloze Task generation:** A physician (MSK) selected a random subset of radiology reports, from which representative
selections of text were extracted and modified to reflect commonly seen radiographic descriptions and findings. A total of 20 cloze prompts were created for use in the survey. Each cloze task prompt was 1-3 sentences in length to provide sufficient context for human and models to identify the masked term. BERT-based models use WordPiece tokenization, which may generate sub-words. Accurate comparison between the CWEs requires that the masked word be present in its complete word form as a token within each model’s vocabulary, i.e. not as a subword. To assess this, each cloze task prompt was tokenized with each of the 3 BERT-based models being studied using the transformers library. Only overlapping complete word forms from each model were selected for masking.

Survey administration: We created and administered a survey of our 281 term pairs and 20 cloze task prompts using the Research Electronic Data Capture (REDCap) tool hosted at VUMC. A convenience sample of 15 healthcare students or professionals were asked to participate in the survey. The instructions provided to participants was adapted from another concept similarity survey, WordSim353. Study data were collected and managed in REDCap.

Human judgment of term pair similarity was assessed using a 7-point Likert scale. For the cloze task portion, survey participants were asked to enter only a single word for each prompt for accurate comparison to BERT-based models because these models have single word vocabularies.

Survey Analysis: Upon completion, survey results were analyzed using R (version 4.0.3). For each term pair, the mean and standard deviation was calculated. Pearson and Spearman correlation coefficient values were calculated with the gensim library using the mean values from our survey and cosine similarity for each NCWE. We also calculate Pearson and Spearman correlation coefficients using the original and modified versions of the UMNSRS-similarity and UMNSRS-relatedness benchmarks introduced above. If either concept in a concept pair was not present in the embedding model vocabulary, that concept pair was not used to calculate the correlation coefficient.

To evaluate CWE models, we composed 20 fill-in-the-blank prompts that reflect text that may appear in a CT chest radiology report, e.g. “Infiltrate in the right middle lobe is seen concerning for ____.” To assess human judgment, survey participants were asked to input free-text for what they determined to be the most likely single word. Survey results were manually reviewed and tabulated to construct a list of the words provided for each cloze task prompt ordered by frequency. Each of our CWE models were then used to also predict the expected word for each cloze task. For word prediction accuracy, we designate the word entered with the highest frequency from our survey as the “expected” word. Accuracy was defined as the percentage of items for which the “expected” word is among the CWE model’s top $k$ predictions for $k \in \{1, 5\}$. We then compared the accuracy for each of our CWEs. We use the transformers library to find the top $k$ predicted words by also providing special tokens as detailed by Goldberg.

Results
Our corpus contained 479,850 documents and comprised a total of 124,892,727 tokens. These documents included other associated clinical documents, including critical result messages. For the NCWEs, our custom models developed using the word2vec and fastText methods each had a vocabulary of 51,002 words, whereas our custom spherical embedding model vocabulary comprised 51,126 words.

Comparison to Fleischner glossary: Using exact matching, the overlap between each of our 3 custom embedding models and words from the Fleischner glossary was 1,207 distinct words. The intersection between the Fleischner glossary vocabulary and BioWordVec, word2vec (Wikipedia) and spherical embedding (Wikipedia) models was 1,272, 1,138 and 1,124 words, respectively. The overlap of all of the embedding model vocabularies with the Fleischner glossary yielded a shared 1,059 words (Figure [1]).

NCWE similarity: The cosine similarity distribution using our term pair selection approach returned an approximately normal distribution for our custom word2vec and fastText models and BioWordVec centered near 0.5. Our custom spherical embedding model and the public models trained using an English Wikipedia corpus were skewed (Figure [2]).

Survey participant characteristics: We had a total of 13 participants who completed the survey. Of these 13, 9 identified as attending physicians. The remaining identified as a medical student (1), resident (1), fellow (1), nurse (1) and advanced practice provider (1). Clinical background for the resident, fellow and attending physicians included anesthesiology (1), emergency medicine (1), family medicine (1), internal medicine (5) and radiology (1). The number of years in practice post-residency for the fellow and attending physicians was a median of 3 years. Participants
reported currently practicing or training in different regions within the US: Northeast (1), Midwest (7), South (4) and West (1).

**Term pair similarity:** The computed Pearson correlation coefficient ($\rho_p$) and Spearman correlation coefficient ($\rho_s$) values comparing mean survey similarity scores to cosine similarity for each model are shown in Table 2. Of each of the 6 models studied, BioWordVec provided the highest $\rho_p$ and $\rho_s$ of 0.60 and 0.61, respectively. Among the embedding models trained using our radiology corpus, the word2vec model yielded the lowest correlation between human and model similarity scores ($\rho_p$, 0.34; $\rho_s$, 0.35). $\rho_p$ and $\rho_s$ values from our custom fastText were 0.40 and 0.41; for our spherical embedding model, $\rho_p$ and $\rho_s$ were calculated as 0.40 and 0.44, respectively. For the English Wikipedia trained word2vec model, $\rho_p$ and $\rho_s$ were 0.31 and 0.32, respectively; the spherical embedding Wikipedia model returned a $\rho_p$ of 0.26 and $\rho_s$ of 0.27.

**Table 2:** Performance of the 6 embedding models on our custom term pair similarity benchmark, RadSim281.

<table>
<thead>
<tr>
<th>Model</th>
<th>Pearson coefficient</th>
<th>Spearman coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>word2vec (ours)</td>
<td>0.34</td>
<td>0.35</td>
</tr>
<tr>
<td>fastText (ours)</td>
<td>0.40</td>
<td>0.41</td>
</tr>
<tr>
<td>spherical emb (ours)</td>
<td>0.40</td>
<td>0.44</td>
</tr>
<tr>
<td>BioWordVec</td>
<td>0.60</td>
<td>0.61</td>
</tr>
<tr>
<td>word2vec (Wiki)</td>
<td>0.31</td>
<td>0.32</td>
</tr>
<tr>
<td>spherical emb (Wiki)</td>
<td>0.26</td>
<td>0.27</td>
</tr>
</tbody>
</table>

**Performance on UMNSRS benchmarks:** $\rho_p$ and $\rho_s$ results from each of our 6 embedding models on the original and modified forms of UMNSRS-Similarity and UMNSRS-Relatedness benchmarks are shown in Table 3. For comparison of word2vec model performance on different training corpora, we present previously published algorithm performance on the modified UMNSRS benchmarks relative to our own findings in Table 4.

**Cloze task accuracy:** Our model pre-trained on a radiology report corpus and BlueBERT each have Top-1 and Top-5 accuracy of 85% and 95%, respectively; the original BERT model had 25% Top-1 and 30% Top-5 accuracy (Table 5).
Table 3: Performance of the 6 embedding models on the original UMNSRS (Pakhomov et al., 2010) and modified UMNSRS (Pakhomov et al., 2016) benchmarks.

<table>
<thead>
<tr>
<th>Model</th>
<th>Similarity</th>
<th>Original UMNSRS</th>
<th>Modified UMNSRS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Pearson</td>
<td>Spearman</td>
</tr>
<tr>
<td>word2vec (ours)</td>
<td>Similarity</td>
<td>0.46</td>
<td>0.44</td>
</tr>
<tr>
<td>fastText (ours)</td>
<td>Similarity</td>
<td>0.52</td>
<td>0.49</td>
</tr>
<tr>
<td>spherical emb (ours)</td>
<td>Similarity</td>
<td>0.35</td>
<td>0.34</td>
</tr>
<tr>
<td>BioWordVvec</td>
<td>Similarity</td>
<td>0.64</td>
<td>0.62</td>
</tr>
<tr>
<td>word2vec (Wiki)</td>
<td>Similarity</td>
<td>0.38</td>
<td>0.38</td>
</tr>
<tr>
<td>spherical emb (Wiki)</td>
<td>Similarity</td>
<td>0.28</td>
<td>0.27</td>
</tr>
<tr>
<td></td>
<td>Relatedness</td>
<td>0.34</td>
<td>0.34</td>
</tr>
<tr>
<td>fastText (ours)</td>
<td>Relatedness</td>
<td>0.41</td>
<td>0.40</td>
</tr>
<tr>
<td>spherical emb (ours)</td>
<td>Relatedness</td>
<td>0.34</td>
<td>0.32</td>
</tr>
<tr>
<td>BioWordVvec</td>
<td>Relatedness</td>
<td>0.57</td>
<td>0.57</td>
</tr>
<tr>
<td>word2vec (Wiki)</td>
<td>Relatedness</td>
<td>0.36</td>
<td>0.35</td>
</tr>
<tr>
<td>spherical emb (Wiki)</td>
<td>Relatedness</td>
<td>0.29</td>
<td>0.27</td>
</tr>
</tbody>
</table>

Note: % OOV is the percentage of out-of-vocabulary words; excluded for correlation calculation.

Table 4: Comparison of word2vec model performance between ours (skip-gram) and Pakhomov et al., 2016 (CBOW) on the modified UMNSRS Similarity and Relatedness benchmark.

<table>
<thead>
<tr>
<th>Corpus Source</th>
<th>Similarity</th>
<th>Relatedness</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pakhomov et al.</td>
<td>Ours</td>
</tr>
<tr>
<td>Clinical text</td>
<td>0.60</td>
<td>0.43</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>0.48</td>
<td>0.39</td>
</tr>
<tr>
<td>PubMed corpus</td>
<td>0.62</td>
<td>NA</td>
</tr>
</tbody>
</table>

Table 5: Top-1 and Top-5 cloze task accuracy for each masked language model.

<table>
<thead>
<tr>
<th>Language Model</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Top-1</td>
</tr>
<tr>
<td>BERT-original</td>
<td>25</td>
</tr>
<tr>
<td>BlueBERT</td>
<td>85</td>
</tr>
<tr>
<td>Our model</td>
<td>85</td>
</tr>
</tbody>
</table>

Discussion
Our term pair creation approach yields a near normal distribution of cosine similarity scores for our custom word2vec and fastText models and the BioWordVvec model. This supports the ability of our approach to capture a normally distributed breadth of similarity scores using word2vec or fastText models trained on biomedical text. We also find that models trained using an English Wikipedia corpus are right skewed in Fig. 2. Additionally, these models consistently performed least well in relation to those models trained using a domain-specific corpus on the RadSim 281, original UMNSRS and modified UMNSRS term similarity benchmarks. These findings suggest poor agreement between models trained using general corpora and clinical judgment.

To our knowledge, our analysis is the first to use spherical embeddings trained using clinical text documents. In contrast to the excellent performance of spherical embeddings reported on general NLP text similarity benchmarks, we find that spherical embeddings often performed less well than other models (Table 2 and 3). The only exception being
that our custom spherical embedding model performed better than our custom word2vec model and at least as well as our custom fastText model on RadSim281. The spherical embedding model proposed by Meng et al. uses directional similarity and jointly learns word and paragraph embeddings. This has been shown to provide improved word similarity and document clustering over other text embedding methods by leveraging both word-word and word-paragraph co-occurrence information. Relative to other text documents, clinical text often contains disparate information in neighboring paragraphs, and this may be an ill-suited substrate for this training strategy. This may cause the model to consider truly dissimilar words as being somewhat similar and may explain the left skew of cosine similarity scores from our spherical embedding model relative to the other 5 models as shown in Fig. 2. Additional study is required to attempt to replicate the benefits of spherical embedding models for document clustering or document classification in the clinical domain.

Among our custom NCWE models, we find that our fastText model performed about as well as our spherical embedding model on RadSim281 (Table 2) and achieves the highest $\rho_p$ and $\rho_s$ on the original and modified UMNSRS benchmarks (Table 3). One of the primary benefits of fastText embedding models is that each word can be represented as a bag of character $n$-grams, which gives these models the ability to handle out-of-vocabulary words. Bojanowski et al. assess the correlation between human judgment and cosine similarity comparing word2vec and fastText models with and without sub-word information using 10 different benchmarks covering 7 different languages. Similar to our own findings, they found that fastText models tend to outperform word2vec models on most term pair similarity benchmarks including the English Rare Word dataset.

We report both Pearson and Spearman correlation coefficients for comparison to earlier studies, some of which provide only the $\rho_p$ and others only the $\rho_s$. The originally proposed BioWordVec fastText model was trained using PubMed and Medical Subject Headings (MeSH). We use the publicly available BioWordVec model trained on PubMed and MIMIC-III corpora. On the UMNSRS benchmarks, we find similar, but slightly lower, $\rho_p$ and $\rho_s$ values compared to the published values (Table 3). Pakhomov et al. evaluate performance on the modified UMNSRS benchmarks using a continuous bag of words (CBOW) word2vec model. Results of their analyses using different training corpora in comparison to our own are shown in Table 2. Our findings for $\rho_s$ using our skip-gram word2vec model and the spherical embedding model trained using an English Wikipedia corpus are much lower on both the UMNSRS-similarity and UMNSRS-relatedness benchmarks compared to their findings (Table 3). Yet, our results from the BioWordVec model are comparable to the $\rho_s$ values they computed using model trained on a PubMed corpus. Although we used the skip-gram training approach and they used the CBOW method, others have found that the skip-gram method provides better or equivalent performance on the UMNSRS benchmarks. This discrepancy is less likely explained by the use of the skip-gram modeling technique instead of CBOW. For models trained on our radiology report corpus or Wikipedia, we were unable to account for one or both concepts for approximately 50% of the modified UMNSRS concept pairs (Table 3). Given that half of these embedding vocabularies are not adequately represented among the concept pairs, our results are likely suffering from bias. This highlights that intrinsic evaluation performance can suffer if a large proportion of the term pair vocabulary is not present within the embedding vocabulary. This discrepancy may also reflect sensitivity of term pair similarity on training parameter selection between the studies. Prior studies show a tradeoff between intrinsic and extrinsic evaluation based upon hyperparameter choice. Chiu et al. find that larger context windows lead to gains in intrinsic evaluation measures with decreased performance on certain downstream tasks. Thus, one must be aware of this and select parameters best suited for the desired objective.

NCWE models can also be augmented with knowledge-based graphs and have been shown to lead to better representations. We attempted to enrich our embedding models using the RadLex ontology, but were met with little success as we found limited RadLex term coverage in our corpus, which is a challenge noted previously. This difficulty may be offset with expansion of the RadLex terminology or by using other existing ontologies that provide improved term coverage.

To our knowledge, this is the first attempt in clinical NLP towards developing intrinsic evaluation benchmarks for CWEs. We find that the original BERT model performs poorly on cloze tasks that reflect radiology report text. We also show equivalent top-1 and top-5 cloze task accuracy between BlueBERT and our custom BERT model (Table 5). These findings appear to suggest that the publicly available BlueBERT model pre-trained using the PubMed and MIMIC-III corpora perform well even in comparison to a model pre-trained using a targeted domain corpus of radiology reports.
from which the cloze task prompts were established. Further study of the generalizability of BlueBERT on other intrinsic evaluation tasks is required. The implication of a generalizable clinical BERT-based model may mitigate the need to pre-train custom BERT-based models for different clinical purposes. This becomes especially important given the sizable financial and environmental costs incurred by language model training. One of the limitations of the BlueBERT model is that it uses the original BERT model vocabulary, rather than training a custom tokenizer using the PubMed and MIMIC-III corpora. BERT-based models can only accept a maximum sequence input length of 512 tokens. Many clinical terms are not present within the original BERT vocabulary and are thus tokenized to subwords for model input. This restricts the number of words that may be used as input for the BlueBERT model, whereas a custom tokenizer would have helped mitigate this issue to some degree.

It remains unclear if a tradeoff between intrinsic and extrinsic evaluation performance exists for CWEs as is reported in NCWEs. Moreover, further study is required to determine if and to what degree the differences in cloze task performance persist after model fine-tuning.

Few, if any, benchmarks exist for intrinsic evaluation of CWE and NCWE models in the clinical domain. Given the potential impact these and other models may have on medical decision making, it is vital to probe models to identify potential flaws and biases. Furthermore, we must be cognizant of existing biases inherent within the clinical data used to train these models. If our models encode biased representations, their deployment may propagate these biases forward and maintain or exacerbate existing healthcare disparities. Probing of models trained on general corpora using intrinsic evaluation methods have unearthed gender, racial and religious biases. With limited model probing tasks within clinical NLP, we contend that this area requires further study especially given the evidence of racial and other disparities present within clinical documentation, management and outcomes. Moreover, while models trained using the MIMIC-III corpus (BioWordVec and BlueBERT) perform admirably, these documents primarily reflect an intensive care patient population and may present an additional source of bias worth consideration.

Among this study’s limitations is that we restricted intrinsic evaluation tasks for NCWE and CWE models to single words. Future studies may use methods to generate multi-word terms while still allowing for comparison within the embedding space. Moreover, our term pair generating method excluded term pairs that were not present in all of the embedding vocabularies. Most terms were excluded because they were not in the NCWE vocabularies trained on a Wikipedia corpus. This design choice may have biased our term pair selection away from relevant medical terms and towards common English words. Another limitation is that our radiology report corpus was primarily limited to CT scans inclusive of the chest. One may find improved results using a corpus of radiology reports from all imaging modalities. Likewise, our corpus was curated from a single academic medical center, which limits the size of our corpus and influences the conditions and findings identified in reports to those found within the region.

Conclusion
Our work introduces two new intrinsic evaluation methods for use among clinical NLP researchers: term pair similarity to compare NCWEs and cloze task accuracy for CWE models. For NCWEs trained on a domain-specific corpus, our results highlight that fastText models tend to outperform word2vec and spherical embedding models. We also demonstrate that the gains afforded by spherical embedding models in general NLP intrinsic evaluation tasks fail to translate to the clinical domain. This emphasizes the need for caution and rigorous evaluation prior to adoption of methods that may excel in general NLP tasks. Importantly, we find that embedding models trained using PubMed and MIMIC-III corpora - BioWordVec and BlueBERT - perform at least as well and often better than models trained on a targeted domain corpus on intrinsic evaluation tasks. This provides additional evidence in support of these biomedical corpora capturing word representations in agreement with clinician judgment. Further study is needed to the establish additional benchmarks and probing tasks. These will help to facilitate language model evaluation to assess for potential biases and determine agreement with clinician judgment. Such analyses are necessary to engender trust and promote understanding of NLP models for broader clinical adoption.

References


[16] BERT Vector Space shows issues with unknown words · Issue #164 · google-research/bert;. Available from: https://github.com/google-research/bert/issues/164.


Machine Learning Predictability of Clinical Next Generation Sequencing for Hematologic Malignancies to Guide High-Value Precision Medicine

Grace Y.E. Kim¹, Morteza Noshad, PhD², Henning Stehr, PhD³, Rebecca Rojansky, MD, PhD³, Dita Gratzinger MD, PhD³, Jean Oak MD, PhD³, Rondeep Brar, MD⁴, David Iberri, MD⁴, Christina Kong, MD³, James Zehnder, MD ³⁴, and Jonathan H. Chen, MD, PhD²⁵

¹Department of Computer Science, Stanford, CA; ²Stanford Center for Biomedical Informatics Research, Stanford, CA; ³Department of Pathology, Stanford, CA; ⁴Department of Hematology, Stanford, CA; ⁵Division of Hospital Medicine, Stanford, CA

Abstract

Advancing diagnostic testing capabilities such as clinical next generation sequencing methods offer the potential to diagnose, risk stratify, and guide specialized treatment, but must be balanced against the escalating costs of healthcare to identify patient cases most likely to benefit from them. Heme-STAMP (Stanford Actionable Mutation Panel for Hematopoietic and Lymphoid Malignancies) is one such next generation sequencing test. Our objective is to assess how well Heme-STAMP pathological variants can be predicted given electronic health records data available at the time of test ordering. The model demonstrated AUROC 0.74 (95% CI: [0.72, 0.76]) with 99% negative predictive value at 6% specificity. A benchmark for comparison is the prevalence of positive results in the dataset at 58.7%. Identifying patients with very low or very high predicted probabilities of finding actionable mutations (positive result) could guide more precise high-value selection of patient cases to test.

Introduction

Next generation sequencing (NGS) has revolutionized research in the biological sciences and has expanded the type of medical care we can provide. NGS based testing has made it possible to detect disorders in their early stages and has opened the gateway towards precision medicine¹. Ideally, such tests can be used frequently for early detection of a disorder and utilized to personalize as much of the disease management process as possible. However, with rising healthcare costs and the already overburdened healthcare system, physicians must and are striving to limit ordering to only when they are at a decision point and believe the outcome of the test will strongly affect the path they choose to go down. However, there is often too little information or too much information to synthesize when the decision has to be made. This is only further exacerbated when these tests are utilized for highly specialized clinical scenarios as is the case with the Heme-STAMP. Heme-STAMP (Stanford Actionable Mutation Panel for Hematopoietic and Lymphoid Malignancies) is a next generation sequencing based test panel. Hematopathologist often utilize Heme-STAMP for diagnostic purposes when there are hematolymphoid process where clinical, histologic, immunophenotypic, and sometimes cytogenic (FISH) information is insufficient to either render a diagnosis of malignancy, or to subtype it in a manner that is useful for sufficiently personalized clinical management. It is also used to map the progression of such disease states and to evaluate if identified laboratory abnormalities (such as cytopenias) are potentially due to hematological malignancies or something else. And is also commonly used to monitor the progression of a disease state (Figure 1). Heme-STAMP uses PCR or hybridization-based DNA capture methods alongside a “targeted sequencing” approach to detect recurrent gene fusions and to screen somatic mutation hotspots of cancer genes².
Given that there are a number of factors considered when determining whether to order a Heme-Stamp test, the objective of this study is to assess how well Heme-Stamp pathological variants can be predicted given electronic health records data. An accurate prediction could add the same level of informational value to care management while saving the cost of running an actual test. Because in many cases the value of the Heme-Stamp test lies in uncovering pathogenic variants, it is likely that a negative prediction will add value to the care management process while reducing the need to have the actual test ordered. A positive prediction by contrast may support the need to order the test in order to identify the specific pathogenic variants and their variant allele frequencies. Given that a negative prediction could result in not ordering the test, an inaccurate prediction might mean that a disease could bypass detection and thus progress. Because of the weight of this consequence, the negative labels were evaluated closely.

Methods

Data. A total of 2,026 Heme-Stamp clinical test results completed between June 2018 and 2020 were used. These samples were drawn from 1,743 Stanford Healthcare patients. Of these patients, 192 patients had multiple tests run over the roughly two-year period. The maximum number of tests run on a single patient was six, but the majority of these multi-test-patients (96%) had four or less tests run. Every sample in the dataset corresponds to a unique specimen. Tests for ctDNA and MRD panels were not included. All specimens underwent a routine sample quality assessment before reporting. Specimens that failed clinical quality control criteria were not included in the dataset.

These Heme-Stamp clinical test results were retrospectively combined with their respective electronic health record data from Stanford Medicine Research Data Repository (STARR)\(^3\). Patient features included prescribed medications, lab values of diagnostic tests, past diagnoses, demographics, and family medical history. Heme-Stamp results were overall categorized as “positive” if a pathogenic mutation was found at a variant allele frequency of \(\geq 5\%\) and “negative” if no pathogenic mutations were found at that variant allele frequency or if the only mutations found were variants of unknown significance (VUS).

The prevalence and rate of negative results of certain diagnoses and sample types were of particular interest. As described earlier, when clinicians order the Heme-Stamp test they may have an existing diagnosis that they want to further subclassify (or check the progression of) or they may have a suspected diagnosis that the Heme-Stamp would be used to verify. Leukemia and Myelodysplastic Syndrome (MDS) are the common hematological malignancies that the clinicians are trying to monitor or evaluate as demonstrated in Figure 2. Respectively, we categorized patient populations into those with a known diagnosis of leukemia (but not of MDS), diagnosis of MDS (but not of leukemia), diagnosis of both MDS and leukemia, or no diagnosis of either disease. Patient history of diagnosis of either disease was identified from free text diagnosis summaries stored in the EHR. Because of the variety of forms and stages of leukemia and MDS, we used regular expressions (such as “%leukemia\%” and “%MDS\%” in various abbreviations and upper/lower case variations) to parse the diagnosis descriptions to best encapsulate patients into their correct disease categorizations. To reflect their relevance in different clinical workflows, we included specimen sample type as a key categorical feature as well.

Predictors. Demographic information, past diagnoses, lab orders, prescribed medications, and family histories were
selected as indicator features that the model could use to predict positive or negative labels. The specific diagnoses, lab orders, prescribed medications, and family histories to use were found by selecting those that were most common among the patients in this study. These features were incorporated by indicating the presence or absence of the listed diagnoses, medications, and family histories using binary labels and by doing simple calculations to include derivatives of lab result values for each selected lab test. These derivatives were found by looking at the results of all past incidences of the specific lab test to find the minimum and maximum result values, the oldest and most recent result values, the average, the sum, and the slope across all the past lab result values. The feature values are summarized in Table 1.

### Table 1. Features/predictors used by the model

<table>
<thead>
<tr>
<th>Past Diagnoses</th>
<th>Leukemia, Lymphoma, Other long-term therapy, Thrombocytopenia, Neoplasm, Skin eruption, Sezary, Hypertension, Myelofibrosis, Mycosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lab Orders</td>
<td>WBC, Hemoglobin, RDW, MCV, Eosinophil, Monocyte, MCH, LDH, CD34, CD3PanT, CD48, CD19, Anion, IgG, Basophils, RBC, Globulin, Lymphocyte, Neutrophil, BUN, eGFR, Albumin, Glucose Serum, Calcium, Creatinine, Alkaline Phosphatase</td>
</tr>
<tr>
<td>Prescribed Medications</td>
<td>Dexamethasone, Ondansetron, Lidocaine, Heparin, Sodium Chloride, Alteplase, Epinephrine</td>
</tr>
<tr>
<td>Family History</td>
<td>Cancer, CAD, Other</td>
</tr>
<tr>
<td>Demographics</td>
<td>Sex, Age</td>
</tr>
<tr>
<td>Derived Features</td>
<td>For each numerical lab test, found patient’s maximum result values, minimum result values, average result values, sum of all result values, most recent lab value, oldest lab value, and slope of all result values</td>
</tr>
</tbody>
</table>

**Model Development and Specification.** The primary model used was an XGBoost (eXtreme Gradient Boosting) classifier, using a gradient boosted decision tree algorithm. This means that the model builds and aggregates the results of multiple learned decision trees in an ensemble approach to make final predictions. Instead of allowing every tree to see the entire dataset, each tree only sees a subset of the training data so that each captures a different signal which can overall be combined to have a more finely tuned model. Additionally, each tree may use a different subset of features. The subset used for each decision tree may be chosen through random selection or based on some sort of metric such as accuracy, Gini index, entropy, etc. that results as a consequence of using different subsets of features. Trees are built sequentially so that each successive tree focuses on properly classifying the observations that were misclassified by the previous tree. The model has various hyperparameters that can be adjusted, prompting us to evaluate multiple combinations: learning rate (0.01, 0.03, 0.1), n_estimators (100, 300, 600, 1000), max_depth (3, 5, 10), and subsample (0.5, 0.75, 0.9). Ultimately the following hyperparameters yielded the best results:

- N_estimators = 300 (number of decision trees used in the ensemble)
- Learning_rate = 0.03 (multiplier for the contribution of each successive classifier in the ensemble)
- Max_depth=5 (maximum depth any decision tree is allowed to have)
- Subsample = 0.5 (fraction sampled from the total dataset to build each tree)

XGBoost and Random Forest both use a decision tree algorithm but differ in how they ensemble the decision trees used. As described earlier in the paper, XGBoost Classifiers build trees sequentially. Random Forest Classifiers on the other hand build trees independently and then average the results of all the trees at the end. To see if the unique ensemble approach of XGBoost contributed to model performance, we compared results to that of Random Forest. We trained a Scikit-Learn Random Forest classifier with the following hyperparameters: n_estimators = 300 and max_depth = 5.

We trained a simple logistic regression model from the Scikit-Learn library in order to compare performance from decision tree-based models to linear models and experimented with sample weights inputted to the XGBoost model. As described earlier in the paper, there is a greater interest in model performance on the negative values so XGBoost models with modified sample weights were tested. Positive values were kept at a weight of 1 but weights of 1.5 and
2.5 for the negative values were experimented with. This difference in weighting would translate to greater importance placed on negative labels during the model’s training/optimization process.

In some of our initial work, we also explored LASSO (a regularization-based model) and Support Vector Machine (a max margin classifier) but found their performance to be lacking so for the rest of the study we moved forward with only the decision tree algorithms and logistic regression model described above.

Model Evaluation. The prediction model was designed to output the pre-test probability of a positive Heme-STAMP result given all the available feature information. If the pre-test probability was greater than a certain threshold, a positive result was predicted, otherwise a negative test was predicted. The accuracy was generated based on a threshold of 0.5 but the AUROC was generated by identifying the threshold that yielded the best balance between the false positive rate and the true positive rate. The false positive rate (FPR) represents the probability that a truly negative result is predicted to be positive and the true positive rate (TPR), also known as recall or sensitivity, presents the probability that a truly positive result is predicted to be positive. The negative predictive value (NPV) and true negative rate (TNR) of these pre-test probabilities at various thresholds were used to generate the TNR vs. NPV graph. The NPV represents the probability that a predicted negative result truly is negative, and the TNR represents the probability that a truly negative result is predicted to be negative (i.e., specificity). The TPR and precision of these pre-test probabilities at different thresholds were also used to generate the Precision-Recall Curve. Precision represents the probability that a predicted positive result is truly positive (i.e., positive predictive value).

We used 10-fold stratified cross validation⁸ with shuffling to train and evaluate the model. For each of the 10 folds, the data was shuffled and then 90% was selected to be part of the training set and 10% to be part of the test set. The data was stratified so that each fold contained a class ratio similar to that of the overall dataset. This was to limit the amount of class imbalance in each fold. The training dataset was used by the model to explore and learn to differentiate between samples with positive and negative labels, indicating identification or lack thereof of a pathogenic variant by Heme-STAMP testing. The test dataset was used to evaluate the performance of the now trained model on new data.

Results

The model was tested on 2,026 Heme-STAMP clinical test results using the 10-fold stratified cross validation method described above. The top 5 features were age, leukemia diagnosis, myelofibrosis diagnosis, sex, and hypertension diagnosis. Table 2 compares the performance of the random baseline and the different models: Logistic Regression, Random Forest, XGBoost (weighted), and XGBoost (unweighted). The random baseline makes only positive predictions and as expected has an accuracy equivalent to the 59% prevalence of positive test results in the dataset. All the models demonstrate accuracy levels greater than the upper limit of the 95% confidence interval for the random baseline.

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
<th>Accuracy 95% C.I.</th>
<th>AUROC</th>
<th>AUROC 95% C.I.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Baseline</td>
<td>59%</td>
<td>48-53%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>60%</td>
<td>58-62%</td>
<td>0.66</td>
<td>0.63-0.69</td>
</tr>
<tr>
<td>Random Forest</td>
<td>62%</td>
<td>60-63%</td>
<td>0.72</td>
<td>0.70-0.74</td>
</tr>
<tr>
<td>XGBoost (unevenly weighted samples)</td>
<td>69%</td>
<td>67-71%</td>
<td>0.74</td>
<td>0.73-0.76</td>
</tr>
<tr>
<td>XGBoost</td>
<td>70%</td>
<td>68-71%</td>
<td>0.74</td>
<td>0.72-0.76</td>
</tr>
</tbody>
</table>

Table 2. Accuracy and AUROC with 95% Confidence Interval (C.I.) for different models and random selection
Figures 3 and 4 respectively demonstrate the Receiver Operating Characteristic Curve and Precision-Recall Curve for the XGBoost model with no weighting (each class is by default weighted the same and thus effectively has no weights).

Among the models experimented, the weighted XGBoost had closest performance to the unweighted XGBoost model (Table 2). The main intention of the weighted XGBoost model was to see if more heavily weighting negative labels would improve the model’s performance on negative samples. So, in Table 3, we compare the NPV/TNR values of the weighted and unweighted XGBoost models to see that the values are nearly identical. The uneven weighting seems to have had no effect on the model’s performance on the negative test results. Because the NPV/TNR results are so similar, only the NPV vs TNR plot for the unweighted XGBoost model is shown in Figure 4.

<table>
<thead>
<tr>
<th>NPV</th>
<th>XGBoost TNR (weighted samples)</th>
<th>XGBoost TNR</th>
</tr>
</thead>
<tbody>
<tr>
<td>90%</td>
<td>14%</td>
<td>13%</td>
</tr>
<tr>
<td>95%</td>
<td>9%</td>
<td>9%</td>
</tr>
<tr>
<td>99%</td>
<td>6%</td>
<td>6%</td>
</tr>
</tbody>
</table>

Table 2. NPV/TNR values of XGBoost weighted and unweighted models

The prevalence and rate of negative results of diagnosis categories and sample types described in the Data section of the paper are shown in Table 3. The unweighted XGBoost model performance on these subgroups is also shown in this table. To evaluate performance, the data samples corresponding to each subgroup was identified for each fold. The AUROC for each subset was found and then averaged over all ten folds to find the average AUROC and the 95% confidence interval.
Table 3. Prevalence of each subgroup in the total dataset (n = # samples), percentage of negative labels in each subgroup, and the AUROC plus the 95% confidence interval for each subgroup.

Figure 5 plots the top twenty features utilized by the unweighted XGBoost model relative to each other. A built-in feature weight function was used to identify the top twenty most important features based on the 300 decision trees fitted by the model (recall that the model was parameterized with n_estimator=300).

Discussion

Results and Implications. When a negative or positive result can be predicted with high accuracy, it is worth considering replacing the actual test with the predicted result since the outcomes are likely similar. And so in theory, the most high-yield diagnostic ordering strategy would be to only order tests where the predicted outcome is not able to be predicted with high accuracy. In the case of the Heme-STAMP, where it is not just the positive/negative outcome but rather the pathogenic variants detected by the test itself that matters, predicted positives should still be ordered as part of high-yield testing even if they’re predicted with high accuracy. This study sought to understand what types of features, models, and subgroups would demonstrate highest performance, especially among the negative labels. We found that the XGBoost model did best with AUROC 0.74 (95% CI: [0.72, 0.76]). Among the various subgroups, model performance was highest on the subgroup with no diagnosis of leukemia or MDS (AUROC 0.74 (95% CI: [0.71, 0.77])) and the subgroup whose specimen type was blood sample (AUROC 0.76 (95% CI: [0.73, 0.80])). And we were able to obtain 99% negative predictive value at 6% specificity.
**Interpretations.** In Table 2 we can see that XGBoost models (weighted and unweighted) boasted marked performance over the accuracy demonstrated by the random baseline. This shows that the model is indeed identifying patterns to make data-driven predictions. While the XGBoost models had higher accuracy than the Random Forest model, the AUROC confidence intervals greatly overlapped. This demonstrates that while XGBoost’s sequential ensemble method improved the overall accuracy of the model, this didn’t necessarily correlate to how well the model was able to differentiate between the positive and negative cases. Interestingly, while the Random Forest model had similar accuracy as the logistic regression model, all three decision-tree-based models (Random Forest, XGBoost weighted, XGBoost unweighted) had higher AUROC values than the logistic regression model did. This suggests that the ensemble tree approach is better than the linear model approach at distinguishing between classes and potentially that it takes an ensemble tree approach particularly with a sequential ensemble method to see accuracy improvement over that of a linear model.

Among the models experimented, the weighted XGBoost had closest performance to the unweighted XGBoost model (Table 2). The main intention of the weighted XGBoost model was to see if more heavily weighting negative labels would improve the model’s performance on negative samples. So, in Table 3, we compare the NPV/TNR values of the weighted and unweighted XGBoost models to see that the values are nearly identical. The uneven weighting seems to have had no effect on the model’s performance on the negative test results. Because the NPV/TNR results are so similar, only the NPV vs TNR plot for the unweighted XGBoost model is shown in Figure 4. Because the XGBoost models had the higher performance among the models but there was minimal difference between the weighted and unweighted XGBoost models, the unweighted XGBoost model was analyzed for the rest of the study.

Among cases where the physician ordered the Heme-STAMP so they could diagnose, subtype, or monitor a malignancy, leukemia and MDS were the most common such malignancies. However, we can see in Table 3 that they compose less than 50% of all the Heme-STAMP cases. Furthermore, the average negative rate across all subgroups with one or both of the malignancies is about 29% which is lower than the negative rate of the subgroup that has neither of those diagnoses. This is equivalent to saying that the average positive rate across all subgroups with one or both of the malignancies is much higher than that of the subgroup with neither diagnosis. This observation matches the clinical scenario because for patients with one or both of the malignancies, it is either known or highly likely that the patient still has the malignancy, and it is often the case that the Heme-STAMP is sought to further characterize the variants driving the malignancy. As described earlier in the paper, our model is focused on making accurate negative predictions, so it is encouraging to see that the model performance on the subgroup that has neither diagnosis and that also has the higher percentage of negative tests is closest to the overall model performance. Model performance on the subgroup with both diagnoses and on the “Other sample” subgroup is highly variable likely because of the small size of the subgroups.

In Figure 5 where we identify the top 20 features, we can see that aside from demographic information such as age and sex, biological markers such as diagnosis of leukemia and myelofibrosis were heavily utilized by the model. Both are diagnoses of hematological malignancies. Additionally, diagnoses of hypertension and thrombocytopenia indicate abnormalities with your blood circulation and blood work. Other hematological malignancies such as lymphoma and other labs related to abnormalities in blood work can be seen among these top 20 features. The model’s incorporation of these features that have clinical relevance provides assurance and credibility to the factors that the model is considering in its prediction.

**Case Reviews.** While we were able to reach high negative predictive values on the tests the model predicted to be negative, qualitative analysis of some of these patients highlighted some reasons why clinicians may still choose to order the test. In one such case, the clinician agreed the patient was very unlikely to have an actionable mutation, but the patient was very anxious and had the financial flexibility to take the test even if it was unlikely to be useful. In another case, the clinician had realized that the patient had a certain baseline mutation and wanted to use the Heme-STAMP to see if there were any other mutations that should be identified and used as a baseline to track changes in the patient’s condition. While some of these nuances may be grappled out free-text notes in the EHR, the complexity required to successfully do so would quickly outweigh the benefits. However, for most cases, a negative prediction from the model can actually lead to the same net information gain as an actual negative test without the cost of it. And even in nuanced cases, such as those listed above, the predicted result can still provide value to the clinician’s thought process and can serve as an additional piece of evidence to support their suggestions as they discuss various potential next step options with patients.
Even as we expand the use of NGS testing and reduce its cost, patients who do not have an established diagnosis, the population types of Next Generation Sequencing tests. Confidence will increase. As the number of Heme-Stamp tests ordered continues to increase, the number of patients for which we can prevent a low probability of having a negative result (i.e., patients with high probability of having a negative result) and thus potential candidates for testing exclusion. We identified patients that have very low probabilities of having a test outcome with an AUROC of 0.74 electronic health records data. We found that by using electronic health records data readily available by the time of the sample, we allowed the model to utilize an ensemble of a high number of estimators. We balanced the cost brought to training time by utilizing XGBoost instead of a different ensemble decision tree model because XGBoost has shown to run with better expediency than other, similar such models.

To mitigate the pitfalls of overfitting, the data was randomly split into a train/test set and the maximum depth of the trees and subsample ratios were kept low. Additionally, early stopping rounds were established so that during the training process the model needed to demonstrate improvement in its evaluation metric (used log loss) every 20 rounds to continue training. This helped to ensure that the model did not continue to fit the model once it had reached a plateau in performance. The literature has shown that implementing early stopping rounds helps to keep overfitting at bay. Due to the limitations in sample size, we allowed the model to utilize an ensemble of a high number of estimators. We balanced the cost brought to training time by utilizing XGBoost instead of a different ensemble decision tree model because XGBoost has shown to run with better expediency than other, similar such models.

Although sample quality and tumor content are routinely assessed before sequencing, the results may not always be accurately determined so the effect of poor sample quality on the number of negative sequencing results cannot be ruled out. However, the fact that the model can predict a subset of negative results with very high accuracy suggests that there are clinical factors characteristic to the patient’s health condition that can be learned from the electronic health record. If the majority of these negative results were due to poor sample quality, there would be no underlying factor that the model could learn from to thus make these predictions with such high accuracy.

Additionally, it is worth noting that algorithmic decision support systems such as the one described here should and can only be used to augment but never to replace clinical judgment. Only a well-trained physician can account for the complex clinical context for each individual patient. In the long-term, a combination of excellent education and data-driven tools will yield the best possible care for the greatest number of patients.

**Conclusion**

The objective of this paper was to see how well Heme-Stamp pathological variants could be predicted given electronic health records data. We found that by using electronic health records data readily available by the time of testing we could predict test outcome with an AUROC of 0.74 (95% CI: [0.72, 0.76]). Furthermore, we were able to identify patients that have very low probabilities of having a positive Heme-Stamp result (i.e., patients with high probability of having a negative result) and thus potential candidates for testing exclusion. These patients accounted for about 6% of all the negative tests but they could be predicted with 99% accuracy. As the number of Heme-Stamp tests ordered continues to increase, the number of patients for which we can prevent a low-yield test with high confidence will increase. This work also demonstrates promising potential to build similar prediction models for other types of Next Generation Sequencing tests. Additionally, because this predictive algorithm can also be used for patients who do not have an established diagnosis, the population it is able to provide value for will continue to grow even as we expand the use of NGS testing and reduce its cost.
Acknowledgements

We would like to thank the Stanford Pathology Department for funding this project through the Value Based Care Initiative and Kathleen Cederlof for her support through this initiative as well. We would also like to thank the Stanford Office of the Vice Provost for Undergraduate Education (VPUE) for funding this project through the VPUE Faculty Grant for Undergraduate Research. Additionally, Jonathan H. Chen’s research is supported by the NIH/National Library of Medicine via Award R56LM013365, the Gordon and Betty Moore Foundation through Grant GBMF8040, the National Science Foundation SPO181514, and the Stanford Clinical Excellence Research Center (CERC). This research used data provided by STARR, STAnford medicine Research data Repository,” a clinical data warehouse containing live Epic data from Stanford Health Care (SHC), the University Healthcare Alliance (UHA) and Packard Children’s Health Alliance (PCHA) clinics and other auxiliary data from Hospital applications such as radiology PACS. The STARR platform is developed and operated by Stanford Medicine Research IT team and is made possible by Stanford School of Medicine Research Office. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or Stanford Healthcare.
References

Predicting Motor Responsiveness to Deep Brain Stimulation with Machine Learning

Kevin J. Krause, BS1, Fenna Phibbs, MD, MPH2, Thomas Davis, MD2, Daniel Fabbri, PhD1

1Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America
2Department of Neurology, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America

Abstract

Deep brain stimulation is a complex movement disorder intervention that requires highly invasive brain surgery. Clinicians struggle to predict how patients will respond to this treatment. To address this problem, we are working toward developing a clinical tool to help neurologists predict deep brain stimulation response. We analyzed a cohort of 105 Parkinson’s patients who underwent deep brain stimulation at Vanderbilt University Medical Center. We developed binary and multicategory models for predicting likelihood of motor symptom reduction after undergoing deep brain stimulation. We compared the performances of our best models to predictions made by neurologist experts in movement disorders. The strongest binary classification model achieved a 10-fold cross validation AUC of 0.90, outperforming the best neurologist predictions (0.56). These results are promising for future clinical applications, though more work is necessary to validate these findings in a larger cohort and taking into consideration broader quality of life outcome measures.

Introduction

Parkinson’s disease (PD) is a relatively common movement disorder and affects 572 per 100,000 people worldwide.1 PD symptoms vary by patient and often include tremor, rigidity, stiffness, and trouble walking.2, 3 Deep brain stimulation (DBS) is a highly complex surgical intervention for PD and other movement disorders, such as essential tremor, and dystonia.2 DBS uses an implanted electric pulse generator to deliver electrical stimulation to specific areas in the brain that control movement.2 Successful electrode implantation and programming is highly complex and requires collaboration from several neurological and neurosurgical disciplines.4-6 Despite undergoing invasive surgery, many DBS patients do not experience an improvement in symptoms4-6 DBS may also worsen symptoms, such as those stemming from dementia7, and lead to adverse reactions, of both minor and major significance.8 Occasionally, additional invasive surgery is required to remove or replant the pulse generator.8

Researchers have identified many factors which affect PD progression and DBS responsiveness. Active lifestyles, caffeine consumption, and moderate alcohol consumption are associated with less severe PD symptoms.9 Conversely, family history, pesticide exposure, rural living, and well water drinking are associated with higher rates of PD onset.10 Interestingly, patients with left-sided symptoms experience less severe motor symptoms, while patients with right-sided symptoms experience less severe cognitive symptoms.11-13 Reduced volume of the brain region called the putamen is also associated with more severe PD progression.14 Strong DBS responsiveness (defined as reduction of symptoms post-DBS) is associated with higher responsiveness to treatment with levodopa, lower baseline tremor severity, and lower age.15 These associations are often contradicted and debated between studies,16-18 a fact which underscores the need for a better understanding of the factors influencing PD and DBS. These challenges and the risks of surgical implantation necessitate a better system for predicting DBS response so that weak responders can be screened out prior to surgery.

Machine learning (ML) is a computational method for identifying patterns in datasets, and has made a large impact in clinical settings from clinical decision support to surgical assistance.19 We are interested in applying ML to improve candidate evaluation and patient counselling prior to DBS. Habets et al. trained a logistic regression ML model to identify strong and weak responders to DBS within a population of 86 PD patients (AUC: 0.79).20 ML has also been used to improve DBS pulse generator programming and electrode placement.21, 22 In this study we build on existing PD, DBS, and ML research to construct a predictive model which distinguishes strong (likely to improve) and weak (unlikely to improve) responders to DBS. We incorporate a wide selection of preoperative variables informed by
clinically known and suspected relationships to PD and DBS. Preoperative variables are chosen to maximize the final model’s applicability to preoperative patient analysis. Finally, we compare our model to clinical specialists to validate its clinical relevance.

<table>
<thead>
<tr>
<th>Demographics</th>
<th>Medical History</th>
<th>Surgery</th>
<th>UPDRS III</th>
</tr>
</thead>
<tbody>
<tr>
<td>Race</td>
<td>Comorbidities: cardiac, thyroid,</td>
<td>Electrode placement locations</td>
<td>Preop on/off-medication</td>
</tr>
<tr>
<td></td>
<td>pulmonary, cancer, neurological,</td>
<td>Electrode settings</td>
<td>Postop on-medication &amp; stim</td>
</tr>
<tr>
<td></td>
<td>and diabetes</td>
<td>Surgery complications / details</td>
<td>Motor Fluctuations (right upper, right lower, left upper, left lower, right total, left total, lip/jaw)</td>
</tr>
<tr>
<td>Age</td>
<td>Alcohol / Smoking / Drugs</td>
<td>Previous surgery details</td>
<td></td>
</tr>
<tr>
<td>Diagnosis</td>
<td>Psychiatric History</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Family History</td>
<td>Provider Information</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Provider Information</td>
<td>Medications</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Year and age of symptom onset</td>
<td>Dopaminergic Drugs &amp; Dosages</td>
<td>Caudate volume (L/R)</td>
<td></td>
</tr>
<tr>
<td>Symptom details and history</td>
<td>Levodopa Daily Equivalent</td>
<td>Thalamus volume (L/R)</td>
<td></td>
</tr>
<tr>
<td>Initial symptom side / location</td>
<td>Anti-depressants &amp; Dosages</td>
<td>Pallidum volume (L/R)</td>
<td></td>
</tr>
<tr>
<td>Current symptom side / location</td>
<td>Anxiolytics &amp; Dosages</td>
<td>Putamen volume (L/R)</td>
<td></td>
</tr>
<tr>
<td>DKEFS</td>
<td>Anti-psychotics &amp; Dosages</td>
<td>Accumbens volume (L/R)</td>
<td></td>
</tr>
<tr>
<td>Fluency</td>
<td>Stimulants &amp; Dosages</td>
<td>Hippocampus volume (L/R)</td>
<td></td>
</tr>
<tr>
<td>Tower</td>
<td></td>
<td>Amygdala volume (L/R)</td>
<td></td>
</tr>
<tr>
<td>Color word naming</td>
<td></td>
<td>WTAR</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>WTAR total score</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Story recall</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Categorized overview of available data. Cells may represent multiple variable fields in the database. Not every variable shown is used in model training. Imaging volumes are reported for left and right areas (L/R).

Methods

Overview and Cohort

We analyzed health records from the Vanderbilt University Medical Center (VUMC) neurocognitive research database. We included a cohort of PD patients who underwent DBS at VUMC. We excluded any patients who had missing on-medication UPDRS III benchmarks before and after surgery. The final supervised-learning cohort contained 105 patients. Patients who did not meet the UPDRS III benchmark requirements for inclusion were reserved in a secondary cohort intended for applying semi-supervised learning. We engineered binary and multcategory target variables to distinguish strong and weak responders to DBS. Last, we trained binary and multcategory machine learning classification models to predict motor improvement. To assess clinical significance, we compared our models’ predictions with those of trained neurologists.

Variables and Data

Table 1 categorizes and summarizes the data available in the VUMC neurocognitive and movement disorders databases. From these databases we extracted preoperative demographic, medical, medication, imaging, and neurocognitive variables. Surgical settings were excluded from analysis, since they do not serve an informative role in deciding to undergo DBS.

This study also examines pre-to-post operative changes in neurocognitive function assessments. The neurocognitive assessments we analyzed include: the Parkinson’s Disease Questionnaire (PD-Q-39)\(^{23}\), the Unified Parkinson’s Disease Rating Scale (UPDRS)\(^{24}\), the Delis-Kaplan Executive Function System (DKEFS)\(^{25}\), the Wechsler Test of Adult Reading (WTAR)\(^{26}\), and the Repeatable Battery for the Assessment of Neuropsychological Status (RBANS)\(^{27}\).

The PD-Q-39 is a 39-question survey which scores quality of life for PD patients.\(^{23}\) The UPDRS is a four-part test which scores non-motor experiences of daily living, motor experiences of daily living, motor examination, and motor complications.\(^{24}\) UPDRS section three (UPDRS III) quantifies motor function and is divided into body subregions.\(^{24}\) UPDRS III also includes body-area subscores, which quantify motor fluctuations in the right upper, right lower, left upper, left lower, right, left, and facial extremities. UPDRS III assessment is performed pre- and post-operation, and on and off medication. DKEFS is a broad set of tests which assess a variety of executive functioning areas, such as verbal fluency. WTAR is a cognitive test for assessing intelligence quotient in traumatic brain injury patients. RBANS assesses cognitive decline in several categories, such as naming, word list recall, and semantic fluency.

Feature Engineering

We defined the age (in years) of symptom onset to be the difference between date of birth and date of symptom onset, which was obtained from the patient during clinical evaluation. We defined pre-operation symptom duration (in years) to be the difference between age pre-operation and age of symptom onset. We defined strong DBS response as improvement (score decrease) in on-medication UPDRS III preop to postop. Weak DBS response is defined as
on-medication UPDRS III worsening (score increase) or staying the same. We did not consider cognitive changes, due to lack of benchmarking data. We also defined a multicategory operation success parameter, using three classes: UPDRS III worsened greater than 25%, UPDRS III improved greater than 20%, and UPDRS III change between 25% worse and 20% better. These class boundaries were chosen to divide the cohort into three evenly sized groups, and to differentiate between high and low magnitude of response with as few classes as possible.

Data Preprocessing and Feature Analysis

We preprocessed our data by imputing missing numeric variables (henceforth features) with each feature’s mean value. Missing boolean and categorical features were labeled with missing-indicators. Categorical features were one-hot encoded into boolean representations of each categorical option. Variance filtering was performed by robust-scaling the feature set and filtering out any features with variance below 0.03. We chose the robust-scaling method, which removes the median and scales to the interquartile range, because it produces boolean and numeric variances on a similar scale for even comparison. Select-K-Best filtering was performed by scaling the feature set to zero mean and unit variance (standard scaling) and choosing the features with the highest DBS-improvement correlation by Pearson’s Chi-squared test (figure 1). The number of features chosen was tuned for performance, and a value of 30 was determined to be optimal. Standard scaling was chosen because chi-squared scores are scaled relative to each feature and we wanted an even weighting of boolean and numeric features in the correlation analysis.

After variance and correlation filtering, we tuned the preprocessing pipeline alongside each predictive model through grid search parameters. The features selected above were fed into the preprocessing pipeline without scaling, so that the pre-classifier scaling method could be tuned as a hyperparameter. We considered standard scaling, minimum-maximum scaling, robust scaling, and no-scaling as preprocessing options. Next, we applied the Synthetic Minority Oversampling Technique to generate 21 synthetic observations of the minority class (weak DBS-response) for an even class balance (63 in each class). The oversampling step’s k-neighbors parameter was tuned within the model grid search. Due to low data quantity (n=105), conjugate undersampling of the majority class was not applied in this study. Last, we considered dimensionality reduction via linear and kernel principal component analysis.

Binary Classification

We trained and compared several supervised and semi-supervised binary classifiers. We trained four model types well studied in the clinical domain, including support vector, logistic regression, k-neighbors, and random forest classifiers. We also trained conjugate semi-supervised classifiers via pseudo-labeling on the reserved semi-supervised cohort. Pseudo-labeling is a technique which uses a trained supervised classifier to predict targets on unlabeled data, so that the newly labeled data can be used to retrain a more generalizable model. The proportion of labelled to pseudo-labeled data used was 10:1; thus 11 additional patients were sampled from the semi-supervised cohort for pseudo-labeling.

Models and preprocessing pipelines were analyzed via grid search with cross validation test scores averaged across 10 stratified hold-out testing folds. Standard deviation across the testing folds was also collected for each metric. Binary metrics included the receiver operating characteristic area under the curve (AUC), accuracy, precision, and recall. The best model was selected to maximize mean cross-validation test AUC.

Multicategory Classification

We trained the same supervised and semi-supervised models as before, but as multiclass predictors. Models were analyzed with cross validation test scores averaged across 10 stratified testing folds. Standard deviation across the testing folds was also collected for each metric. AUC is poorly defined in multiclass problems, so we instead relied primarily on f1-score and mean-squared-error to evaluate the multiclass models. We also evaluated precision and recall scores. F1, precision, and recall scores were macro-averaged across the three classes. The best model was selected to maximize mean f1-macro.

Clinical Performance Comparison

With the goal of deploying a clinical model to reduce the risk of poor outcomes in DBS candidates, we compared the performances of our best models with trained movement disorder experts. Two board certified neurologists reviewed the patients and training features analyzed in this study to predict whether each patient’s motor function would improve after DBS. The neurologists made both a binary (will they improve: yes/no) and multicategory (in which range will they improve) prediction for each patient. The most accurate predictions from both neurologists were combined to form a neurologist-best-case prediction set.
We compared binary AUC, accuracy, recall, precision, and kappa scores, along with multiclass f1, mean squared error, precision, and recall scores. Metrics were averaged across 10 stratified held-out cross validation test sets. Multiclass metrics were macro-averaged between classes.

Additional Statistical Analyses

Two board-certified neurologists reviewed the results of our analyses to identify features of interest for further study. Features of interest were explored with two-sample independent t-tests, allowing unequal variances to make as few assumptions about the data as possible. From these t-tests we obtained 95% confidence intervals for the true differences in mean measures between groups, as well as p-value estimates of statistical significance.

Results

Summary Statistics

1,893 (64.5%) of 2,935 total patient records in our database were diagnosed with PD. Of those with PD, 105 (3.58%) patients had records of both pre- and post-operative UPDRS III scores. Table 2 summarizes demographic data of our final cohort. Most study participants were male (69.5%) and white (91.4%). Black, Hispanic, and Asian participants made up 4.8% of study participants. 27.6% of the cohort were prescribed anxiolytics, and 38.1% of the cohort were prescribed anti-depressants.

Feature Analysis

Figure 1 depicts the top-30 features by Person’s chi-squared correlation with motor improvement. Preop total and extremity tremor scores were highly correlated with DBS response. Also highly correlated were left putamen volume, symptom side, race, and histories of cardiac, psychiatric, and diabetic complications (figure 1).

Table 3 describes the differences in variable ranges between the weak and strong response groups. Strong DBS responders were lower in median levodopa equivalent prescription range, but were higher in median left putamen volume, UPDRS III (on and off) scores, and WTAR scores (table 3, figure 2).

<table>
<thead>
<tr>
<th>Category</th>
<th>Count (%)</th>
<th>Category</th>
<th>Count (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anti-depressants</td>
<td>30 (38.1%)</td>
<td>Anxiolytics</td>
<td>29 (27.6%)</td>
</tr>
<tr>
<td>Citalopram (Celexa)</td>
<td>12 (11.4%)</td>
<td>Clonazepam (Clonopin)</td>
<td>21 (20.0%)</td>
</tr>
<tr>
<td>Sertraline (Zoloft)</td>
<td>7 (6.7%)</td>
<td>Alprazolam (Xanax)</td>
<td>4 (3.8%)</td>
</tr>
<tr>
<td>Duloxetine (Cymbalta)</td>
<td>5 (4.8%)</td>
<td>Lorazepam (Ativan)</td>
<td>3 (2.9%)</td>
</tr>
<tr>
<td>Bupropion (Wellbutrin)</td>
<td>4 (3.8%)</td>
<td>Diazepam (Valium)</td>
<td>1 (&lt; 1%)</td>
</tr>
<tr>
<td>Paroxetine (Paxil)</td>
<td>4 (3.8%)</td>
<td>Race</td>
<td>Count (%)</td>
</tr>
<tr>
<td>Venlafaxine (Effexor)</td>
<td>3 (2.9%)</td>
<td>White</td>
<td>96 (91.4%)</td>
</tr>
<tr>
<td>Sex</td>
<td>Count (%)</td>
<td>Black</td>
<td>3 (2.9%)</td>
</tr>
<tr>
<td>Male</td>
<td>73 (69.5%)</td>
<td>Hispanic</td>
<td>1 (&lt; 1%)</td>
</tr>
<tr>
<td>Female</td>
<td>32 (30.5%)</td>
<td>Asian</td>
<td>1 (&lt; 1%)</td>
</tr>
</tbody>
</table>

Table 2: Select summary statistics of count and percent of final cohort.

Figure 1: Bar graph representing the top-30 features, ranked by chi-squared correlation with improved UPDRS III score. The column variance is listed above each column. Features with zero variance are ignored. The white column text shows the calculated p value for the feature.
Independent two-sample t-tests allowing unequal variances found statistically significant differences in mean left putamen volume, UPDRS III off-medication score, UPDRS III on-medication score, and WTAR score between weak and strong DBS responders (p<0.01, 0.05, 0.01, 0.01, respectively). From these tests we are 95% confident that the true difference in feature means lay between 244 and 926 mm$^3$, 0.4 and 8.4, 3.8 and 11.0, and 0.0 and 12.4, respectively, between response groups. Figure 2 shows a boxplot comparing left putamen volume between DBS response groups.

Figure 3 shows boxplot comparisons of UPDRS III change post-DBS vs race, symptom side, and psychiatric status. All three African American patients showed weak responses to DBS. A two-sample t-test allowing unequal variances found a statistically significant difference in mean UPDRS III improvement for white and black patients (p<0.01). From this test we are 95% confidence that the true difference in mean UPDRS III improvement between these groups lays between 3.9 and 9.9 (white group with greater improvement).

Patients with left-sided symptoms averaged higher score improvements than those with right sided symptoms. A two-sample t-test allowing unequal variances found a statistically significant difference in mean UPDRS III improvement for these groups (p<0.02). From this test we are 95% confidence that the true difference in mean UPDRS III improvement between these groups lays between 2.9 and 20.0 (left-sided group with greater improvement).

### Table 3: Comparison of medians and interquartile ranges (IQR) of select features relative to strong and weak response groups. P values were obtained from two-sample t tests allowing unequal variances. Strong response is defined as a negative change (improvement) in UPDRS III score post-DBS. Weak response is defined as a positive or no change (worsening) in UPDRS III score post-DBS.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Weak Responder Median (IQR)</th>
<th>Strong Responder Median (IQR)</th>
<th>P Val</th>
</tr>
</thead>
<tbody>
<tr>
<td>Levodopa Total (mg)</td>
<td>1328 (954–1752)</td>
<td>1250 (831–1439)</td>
<td>&lt;0.3</td>
</tr>
<tr>
<td>CT Derived Brain Region Volumes (mm$^3$)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Left Putamen</td>
<td>5042 (4547–5634)</td>
<td>5581 (5383–5975)</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>Right Putamen</td>
<td>4984 (4520–5788)</td>
<td>5354 (5032–5740)</td>
<td>&lt;0.3</td>
</tr>
<tr>
<td>Left Accumbens</td>
<td>582 (527–690)</td>
<td>588 (533–694)</td>
<td>&lt;0.9</td>
</tr>
<tr>
<td>Right Accumbens</td>
<td>572 (498–647)</td>
<td>609 (524–678)</td>
<td>&lt;0.2</td>
</tr>
<tr>
<td>Left Amygdala</td>
<td>1356 (1185–1542)</td>
<td>1422 (1261–1700)</td>
<td>&lt;0.2</td>
</tr>
<tr>
<td>Right Amygdala</td>
<td>1613 (1412–1716)</td>
<td>1663 (1434–1869)</td>
<td>&lt;0.5</td>
</tr>
<tr>
<td>Neurocognitive Assessment Scores (score units)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UPDRS III off-meds</td>
<td>37 (32–44)</td>
<td>42 (36–52)</td>
<td>&lt;0.05</td>
</tr>
<tr>
<td>UPDRS III on-meds</td>
<td>16 (11–22)</td>
<td>23 (18–28)</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>WTAR</td>
<td>32 (21–36)</td>
<td>37 (28–46)</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>RBANS WL Recog.</td>
<td>19 (18–20)</td>
<td>19 (18–20)</td>
<td>&lt;0.15</td>
</tr>
<tr>
<td>DKEFS SvSF</td>
<td>-1 (-4–2)</td>
<td>0 (-2–3)</td>
<td>&lt;0.15</td>
</tr>
</tbody>
</table>

**Table 3**: Comparison of medians and interquartile ranges (IQR) of select features relative to strong and weak response groups. P values were obtained from two-sample t tests allowing unequal variances. Strong response is defined as a negative change (improvement) in UPDRS III score post-DBS. Weak response is defined as a positive or no change (worsening) in UPDRS III score post-DBS.

**Figure 2**: Boxplot of DBS outcome vs size of left putamen area in brain. Strong response is UPDRS III improvement (score decrease). Weak response is UPDRS III worsening (score increase).

**Figure 3**: Comparison of UPDRS III score improvement after DBS, grouped by race, symptom side, and psychiatric status. Positive improvements indicate strong DBS response. Negative improvements indicate weak DBS response.
Patients with untreated psychiatric conditions averaged lower improvements in UPDRS III score than those with treated psychiatric conditions. From our two-sample t-test we are 95% confident that the true difference in mean UPDRS III improvement between these groups lays between 1.1 and 10.6. However, we did not find sufficient evidence that there is a difference between these groups (p<0.2).

**Classification Results**

Table 4 shows the cross-validated performance metrics for our best optimized binary and multicategory models. The highest AUC binary model was a supervised support vector classifier (AUC 0.90). The multicategory model with the highest performance was a logistic regression with pseudo-labeling (f1-macro: 50.9, MSE: 0.68). Table 4 also compares our best performing models against the performance of the neurologists. The binary and multicategory models both outperformed the neurologists and neurologists-best-case (best-neurologist AUC: 0.560, best-neurologist f1-macro: 0.264). Table 5 compares the Cohen kappa agreement scores between the best binary model, neurologist A, neurologist B, and neurologist-best-case.

**Discussion**

In this study, we built, optimized, and analyzed eight machine learning models. Logistic regression and support vector classifiers seemed to produce the best fits to our data. Our models outperformed the clinical experts in a variety of performance metrics, suggesting that there may be a viable future for this type of tool in the clinic. The addition of imaging features in our analysis may have produced a significant performance improvement over similar studies which did not include them. In addition, we found several interesting outcome-correlations, including race, symptom-sidedness, psychiatric status, and putamen volume.
Our strongest binary classification model outperformed the trained neurologists. The strongest multiclassification classification model achieved lower performance, yet still outperformed the neurologists. In both cases, the neurologists achieved high recall scores, but low AUC and f1 scores, suggesting a tendency to overpredict strong DBS response (predicting ‘strong response’ for every patient in the binary case yields 100% recall). The low kappa interrater agreement scores between neurologists underscore the clinical difficulty in predicting DBS response. One possible explanation for the discrepancies between the neurologists is that they would not typically make predictions based on a binary indication of change in UPDRS III. Rather, neurologists consider a host of factors around motor fluctuations, medication responsiveness, and quality of life. Our study used a best-case agreement scheme to generate a set of the strongest predictions from our two board-certified neurologist-reviewers. Future analyses may benefit from an additional expert reviewer, as an odd number of reviewers would permit a majority-voting scheme to generate the best-case predictions.

Our best-fit binary model’s results rival those of previous ML studies in this arena. One major difference in our approach is the inclusion of imaging data into our predictor. Despite the performance improvement, imaging features are more difficult to obtain, and future end-users may prefer a tool which does not require them. These findings highlight the challenge of balancing performance with user-friendliness.

This study found a significant positive correlation between larger left putamen volumes and positive DBS response. Interestingly, other studies have noted decreased size and grey matter volume of the putamen in Parkinson’s and other neurodegenerative disorders. Further, research has shown that targeting the putamen with DBS improves motor fluctuations. It is possible that a larger putamen is easier to target with DBS, or that patients with larger putamen regions have less severe symptoms. The exact link is between DBS, tremor, and the putamen, is unknown but the limited research available suggests that a connection is probable. The differences observed in left and right putamen size significances may be due to differences in symptom-sidedness and/or surgical implant sidedness.

There was a strong correlation between race and DBS response. All three African American patients in this study experienced worsened motor symptoms. Additionally, one out of one Asian participant experienced worsened tremors. The small number of nonwhites in the study (5) makes it challenging to accurately analyze the impact of race on DBS response. Other research has indicated a lower prevalence of PD in African-Americans and Latinos, as compared with whites. These differences are believed not to be related to age, sex, income, insurance, or healthcare utilization, but rather biological or other differences. More longitudinal data are needed to thoroughly explore the interesting relationships between race and DBS response.

The difference in symptom sidedness and outcome may be related to differences observed in PD progression between symptom sidedness groups. Many studies have noted less severe cognitive symptoms in patients with right-sided symptoms and less severe motor symptoms in patients with left-sided symptoms. Our study noted motor outcome favorability for left-sided patients, which is consistent with the general observation of less severe motor outcomes in left-sided PD patients.

One limitation of our study is that we only included motor symptom scores (UPDRS III) as success benchmarks, while other studies have included broader features related to quality of life. For example, Habets et al. defined strong response relative to changes in UPDRS II, III, and IV. UPDRS II measures PD difficulties in daily life and UPDRS IV measures complications of therapy, while UPDRS III only measures motor fluctuation severity. Management of motor fluctuations is a major motivating factor in choosing to undergo DBS, so changes in their severity are a reasonable benchmark for DBS response. Further, we were able to gather a larger training cohort by only requiring UPDRS III measurements, which are more regularly recorded in our database. If we had access to more data, we would have liked to have included benchmarks from the PD-Q-39, UPDRS II, or UPDRS IV, which measure factors beyond motor symptoms in PD patients.
Another limitation of this study is the lack of an external validation test set. We chose not to withhold a final test set due to the small size of our dataset (N=105) which could easily become overfit if the sample size were further reduced. We validated our results with 10-fold stratified cross validation, providing an optimistic estimate of our model’s fit.

Conclusion

Our predictive model produced a clinically significant performance improvement. These results are very promising for the future of DBS candidate evaluation, counselling, and expectation-setting. More work is necessary to validate these findings in a larger cohort and taking into consideration broader quality of life outcome measures. However, if these models can be further refined and validated in larger cohorts, it may be possible to deploy such a tool in the clinical setting to better support DBS candidate counselling, evaluation, and expectation setting.

References


Trust and credibility of information sources related to COVID-19 among high-risk ethnically diverse adults at the onset of the New York City outbreak: A cross-sectional survey conducted via a community health portal

Rita Kukafka, DrPH, MA, FACMI¹, Mari Millery, PhD¹, Samuel Pan, MS², Thomas B. Silverman, MPH¹, Julia E. McGuinness, MD², Katherine D. Crew, MD²

¹Department of Biomedical Informatics, Columbia University, New York, NY, USA
²Herbert Irving Comprehensive Cancer Center, Columbia University, New York, NY, USA

Abstract

In March 2020, days after New York shut down to mitigate the spread of COVID-19, we developed a cross-sectional, participant-administered electronic survey to explore how New Yorkers were impacted by and were responding to the ongoing crisis. A critical component of the survey was to assess how credible and trustworthy respondents found various information sources. To advertise and distribute the survey, we embedded an invitation to participate using a popup on the GetHealthyHeights.org website. GetHealthyHeights was designed using community-based participatory research for the medically-underserved, urban, and largely Latinx community of Washington Heights-Inwood, New York City. We received 321 responses from April through July 2020. Participant ages ranged from 25 to 87, and 25% were Latinx. Results showed that the choice of and trust in different COVID-19 information sources were observed to be significantly different across demographic variables, including gender, age, race, and chronic health conditions. In the domains of trust and information source credibility, designers should account for perspectives of diverse subgroups.

Introduction

Beginning in late 2019, a severe acute respiratory syndrome (COVID-19) began to rapidly spread across the globe, becoming an unprecedented public health crisis. While the COVID-19 impact was global, New York City (NYC) quickly became the pandemic’s epicenter. By mid-March, infection rates were five times higher than the rest of the county, with cases one-third of total confirmed US cases.¹ In an effort to contain the spread of COVID-19, the governor's office implemented extraordinary restrictions on businesses, schools, social gatherings, and use of public transportation. These restrictions were fundamentally the same behaviorally focused non-pharmaceutical interventions (NPIs) put into place during the 1918 pandemic ² ³, but unprecedented for the ethnic and diverse populations living and working in the NYC area with access to a multitude of information sources, the Internet and social media.

The novel COVID-19 pandemic has shone a spotlight on the importance of communication and, in particular the critical importance of citizen trust in the communicator. A review of 26 papers found that having a high level of trust in authorities and with communications received about the disease is associated with adherence to behaviors that could not be explained in terms of perceptions of risk, disease severity or coping appraisals.⁴ The uncertainty and constantly evolving science on COVID-19 has made health communication during the pandemic difficult, as it has resulted not only in a massive flow of health information, but also in rapidly changing, mixed and inconsistent messages.⁵ Trust in information sources has been shown to influence perceived severity, increase willingness to adopt behaviors such as physical distancing, and information seeking.⁶ ⁵, ⁷-⁹ Lack of trust in information sources has also been related to increased COVID-19 mortality.¹⁰ Patterns of anxiety resulting in compulsive behaviors such as binge shopping and hoarding have also been associated with lack of trust in information sources.¹¹

In this paper, we describe insights from a time sensitive cross-sectional survey conducted in April 2020 among high-risk ethnically diverse adults in NYC. A critical component of the survey was to assess how credibility and trustworthy respondents found various information sources of information. We present these findings here, as well as how credibility and trust were perceived differently by varying socio-demographic population characteristics.
Methods

Design and setting

In March 2020, days after New York shut down to mitigate the spread of COVID-19, we developed a cross-sectional, participant-administered electronic survey to explore how New Yorkers were impacted by and were responding to the ongoing crisis. Because the survey was intended to capture participants’ perceptions and experiences during the rapidly evolving emergency, we employed an open, voluntary-response sampling frame, which allowed us to administer the survey as quickly as possible. To advertise and distribute the survey, we embedded an invitation to participate on the GetHealthyHeights.org website. GetHealthyHeights was designed using community-based participatory research for the medically underserved, urban, and largely Latinx community of Washington Heights-Inwood, New York City (WAHI).12, 13 A pop-up on the website was added on April 9th, 2020, inviting respondents to share how the outbreak has affected their health and well-being. In addition, we emailed participants engaged in three community-based research studies on breast cancer prevention and the Database Shared Resource at Columbia University Irving Medical Center (CUIMC), inviting them to participate in the survey. This solicitation to CUIMC research participants was conducted to gain survey responses from community members at higher risk for COVID-19 complications due to co-morbidities compared to the general population. Finally, we emailed the listserv of the Columbia Community Partnership for Health, Columbia University’s community outreach organization that connects members of the WAHI community to health resources and research, with an invitation to participate in the survey. The survey took approximately fifteen minutes to complete, and respondents were not compensated for participating. The CUIMC Institutional Review Board approved the study procedures. Data were collected from April through July 2020.

Questionnaire

Sociocultural, healthcare utilization and demographic factors

We selected questions from the validated Flu Telephone Survey Template (FluTEST), which was designed to assess perceptions and behavior during an influenza pandemic.14 Information collected on demographics and background included age, sex, gender, racial/ethnic background, employment, and current medical conditions that could increase risk for susceptibility to and/or complications from COVID-19. In addition, we asked whether participants had been tested, were told by a healthcare provider or believed they had been infected with COVID-19, whether access to healthcare was interrupted, and whether they lost a job due to the pandemic.

Trust in information sources

We assessed credibility of information sources and trust in official agencies with 16 items. Participants were instructed to ‘indicate how credible you believe the below communicators are by selecting the options that best reflect your opinions’ about sources coming from the Federal Government (like the president or vice president), information coming from your governor (for example, Andrew Cuomo (New York) or Phil Murphy (New Jersey)), information coming from your mayor (for example, Bill de Blasio (New York City)) and information coming from scientists. For each source, participants were asked to indicate the degree to which the information source could be trusted, is accurate, tells the whole story, and is biased or one-sided. The responses included probably true, probably false, not sure and no opinion. Participants were also asked their opinions on the various sources from which they received information on COVID-19, including people they speak to daily, healthcare professionals (for example, your doctor, nurse, or pharmacist), official websites (like the Centers for Disease Control or the health department), other websites, social media, and television. Participants rated these sources with response options that included very helpful, somewhat helpful, somewhat unhelpful, very unhelpful and not sure.
Statistical Methods

Chi square tests (and fisher’s exact tests where appropriate) and students t-tests (or non-parametric alternatives such as the Wilcoxon rank-sum test where appropriate) were performed to test for any differences by co-morbidities, age, race, ethnicity and employment status. A p-value of less than 0.05 was considered to indicate statistical significance. Data analyses were performed using SAS 9.4 (SAS Institute).

Results

Participant Characteristics

Between April and July 2020, 321 people responded to the survey. Table 1 summarizes participant characteristics.

<table>
<thead>
<tr>
<th>Table 1: Participant characteristics (N=321).</th>
<th>M (Range) or N (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years)</td>
<td>62 (25, 87)</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>51 (18%)</td>
</tr>
<tr>
<td>Female</td>
<td>225 (82%)</td>
</tr>
<tr>
<td>Race*</td>
<td></td>
</tr>
<tr>
<td>Asian</td>
<td>8 (3%)</td>
</tr>
<tr>
<td>Black or African American</td>
<td>22 (9%)</td>
</tr>
<tr>
<td>Native American or Alaska Native</td>
<td>2 (1%)</td>
</tr>
<tr>
<td>White</td>
<td>207 (85%)</td>
</tr>
<tr>
<td>Multiple</td>
<td>6 (2%)</td>
</tr>
<tr>
<td>Ethnicity*</td>
<td></td>
</tr>
<tr>
<td>Hispanic or Latino</td>
<td>68 (25%)</td>
</tr>
<tr>
<td>Not Hispanic or Latino</td>
<td>202 (75%)</td>
</tr>
<tr>
<td>Geographic Area</td>
<td></td>
</tr>
<tr>
<td>NYC or surrounding commutable areas</td>
<td>246 (91%)</td>
</tr>
<tr>
<td>Other regions of the US</td>
<td>24 (9%)</td>
</tr>
<tr>
<td>Employment</td>
<td></td>
</tr>
<tr>
<td>Working Full Time</td>
<td>102 (37%)</td>
</tr>
<tr>
<td>Working Part Time</td>
<td>37 (13.5%)</td>
</tr>
<tr>
<td>Not Working</td>
<td>37 (13.5%)</td>
</tr>
<tr>
<td>Retired</td>
<td>99 (36%)</td>
</tr>
<tr>
<td>Lost Job/Had Hours Reduced Due To COVID-19</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>42 (16%)</td>
</tr>
<tr>
<td>No</td>
<td>152 (55%)</td>
</tr>
<tr>
<td>Respondent Believes S/he was Infected with COVID-19</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>43 (15%)</td>
</tr>
<tr>
<td>No</td>
<td>174 (58%)</td>
</tr>
<tr>
<td>Don’t Know</td>
<td>81 (27%)</td>
</tr>
<tr>
<td>Tested for COVID-19</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>30 (10%)</td>
</tr>
<tr>
<td>No</td>
<td>266 (90%)</td>
</tr>
<tr>
<td>Respondent Told by Health Care Provider that S/he Likely Has COVID-19</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>22 (7%)</td>
</tr>
<tr>
<td>No</td>
<td>273 (92%)</td>
</tr>
<tr>
<td>Chronic Conditions</td>
<td></td>
</tr>
<tr>
<td>Breathing Disorders</td>
<td>56 (17%)</td>
</tr>
<tr>
<td>Cancer</td>
<td>94 (29%)</td>
</tr>
<tr>
<td>Diabetes</td>
<td>21 (7%)</td>
</tr>
<tr>
<td>Kidney Disease</td>
<td>11 (3%)</td>
</tr>
<tr>
<td>Mental Illness</td>
<td>58 (18%)</td>
</tr>
<tr>
<td>Reported as “other”, not specified</td>
<td>90 (28%)</td>
</tr>
</tbody>
</table>

*To explore the impact of minority status in subsequent analyses, we dichotomized these variables into White/non-Hispanic (N=182, 66.91%) and Other to include other than White/non-Hispanic minority groups (N=90, 33.09%). Missing=49
Respondents ranged in age from 25 to 87 years old, with a mean age of 62. Twenty-five percent (25%) were Hispanic or Latino, and 82% were female. Around thirteen percent (13.5%) were not working and 36% were retired. Ninety one percent (91%) of lived in NYC or in areas commutable to NYC. All respondents reported having at least one chronic condition: 29% reported having cancer, 18% reporting mental illness and 17% reporting breathing disorders. Given that COVID-19 testing was mostly unavailable early in the pandemic, only 10% report that they were tested.

Sociodemographic Factors and Co-morbid Conditions Associated with Sources of COVID-19 Information Sources

Perceived helpfulness of information sources by sociodemographic and chronic health conditions is shown in Table 2. White/non-Hispanic participants were significantly more likely to rank ‘people I speak to daily’ as helpful compared to others. Younger participants, aged 18-45, ranked social media significantly more useful compared with older participants, whereas participants >65 ranked television as more useful compared with younger participants. Females were significantly more likely than males to rank social media and official websites and social media as helpful. With respect to chronic health conditions, respondents reporting conditions other than cancer, mental illness, breathing disorders or diabetes were significantly more likely to rank official websites more useful, and also more likely to be unsure of the usefulness of websites compared with respondents not reporting in this category.

Table 2. Sociodemographic and co-morbid factors associated with COVID-19 information source.

<table>
<thead>
<tr>
<th>Please indicate how helpful you find each of the below sources of information.</th>
<th>People I Speak to Daily</th>
<th>Healthcare Professionals</th>
<th>Official Websites</th>
<th>Social Media</th>
<th>Television</th>
</tr>
</thead>
<tbody>
<tr>
<td>Race</td>
<td>Helpful</td>
<td>Unhelpful</td>
<td>p</td>
<td>Helpful</td>
<td>Unhelpful</td>
</tr>
<tr>
<td>White, non-Hispanic</td>
<td>152</td>
<td>28</td>
<td>0.04*</td>
<td>127</td>
<td>46</td>
</tr>
<tr>
<td>Hispanic</td>
<td>184</td>
<td>30</td>
<td></td>
<td>115</td>
<td>40</td>
</tr>
<tr>
<td>Other</td>
<td>66</td>
<td>23</td>
<td></td>
<td>63</td>
<td>27</td>
</tr>
<tr>
<td>Age Group</td>
<td>18-45</td>
<td>43</td>
<td>0.15</td>
<td>40</td>
<td>17</td>
</tr>
<tr>
<td>46-64</td>
<td>20</td>
<td>0.57</td>
<td></td>
<td>74</td>
<td>26</td>
</tr>
<tr>
<td>65+</td>
<td>100</td>
<td>17</td>
<td></td>
<td>82</td>
<td>32</td>
</tr>
<tr>
<td>Gender</td>
<td>Male</td>
<td>99</td>
<td>32</td>
<td>0.38</td>
<td>38</td>
</tr>
<tr>
<td>Female</td>
<td>103</td>
<td>39</td>
<td></td>
<td>155</td>
<td>61</td>
</tr>
<tr>
<td>Working</td>
<td>Yes</td>
<td>111</td>
<td>25</td>
<td>0.81</td>
<td>98</td>
</tr>
<tr>
<td>No</td>
<td>109</td>
<td>26</td>
<td></td>
<td>95</td>
<td>33</td>
</tr>
<tr>
<td>Lost Job</td>
<td>Yes</td>
<td>28</td>
<td>4</td>
<td>0.01*</td>
<td>26</td>
</tr>
<tr>
<td>No</td>
<td>131</td>
<td>21</td>
<td></td>
<td>110</td>
<td>41</td>
</tr>
<tr>
<td>Cancer</td>
<td>Yes</td>
<td>73</td>
<td>3</td>
<td>0.31</td>
<td>67</td>
</tr>
<tr>
<td>No</td>
<td>154</td>
<td>39</td>
<td></td>
<td>129</td>
<td>58</td>
</tr>
<tr>
<td>Diabetes</td>
<td>Yes</td>
<td>17</td>
<td>4</td>
<td>0.96</td>
<td>16</td>
</tr>
<tr>
<td>No</td>
<td>210</td>
<td>48</td>
<td></td>
<td>180</td>
<td>45</td>
</tr>
<tr>
<td>Mental Illness</td>
<td>Yes</td>
<td>57</td>
<td>9</td>
<td>0.58</td>
<td>41</td>
</tr>
<tr>
<td>No</td>
<td>180</td>
<td>43</td>
<td></td>
<td>155</td>
<td>62</td>
</tr>
</tbody>
</table>

663
Table 2. Sociodemographic and co-morbid factors associated with COVID-19 information source. (Continued)

<table>
<thead>
<tr>
<th>Breathing Disorder</th>
<th>Yes</th>
<th>45</th>
<th>(83.33%)</th>
<th>9</th>
<th>0.68</th>
<th>40</th>
<th>1</th>
<th>0.68</th>
<th>43</th>
<th>11</th>
<th>0.64</th>
<th>23</th>
<th>29</th>
<th>0.38</th>
<th>40</th>
<th>13</th>
<th>0.77</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td></td>
<td>182</td>
<td>(80.80%)</td>
<td>43</td>
<td>19.11%</td>
<td>156</td>
<td>23</td>
<td>0.45</td>
<td>67</td>
<td>23</td>
<td>0.03*</td>
<td>26</td>
<td>56</td>
<td>0.10</td>
<td>62</td>
<td>28</td>
<td>0.19</td>
</tr>
</tbody>
</table>

| Other Chronic Condition | Yes | 73 | (80.22%) | 18 | 0.73 | 68 | 23 | 0.45 | 67 | 23 | 0.03* | 26 | 56 | 0.10 | 62 | 28 | 0.19 |
| No                 |     | 154| (81.91%) | 34 | 18.09%| 128| 54 | 0.33% | 88 | 41 | 14.59%| 42 | 22 | 57.78%| 56 | 43 | 23.63%

Sociodemographic Factors and Co-morbid Conditions Associated with Trust of Information Sources of COVID-19

Participants’ perceptions of the credibility of information coming from the federal government and scientists are shown in tables 3 and 4. White/non-Hispanic participants valued information coming from the federal government and scientists differently than minority respondents. Few among both White/non-Hispanic and minority respondents responded true when asked if information coming from the federal government can be trusted (14.92% and 16.67% respectively). White/non-Hispanic respondents were significantly more likely to respond false when asked if information coming from the federal government can be trusted compared with minority groups (72.93% vs. 47.78% respectively). White/non-Hispanic respondents were also more likely to indicate that information coming from the federal government is inaccurate, biased, and does not tell the whole story. Minority groups were more likely to be unsure or have no opinion about the trustworthiness, accuracy, bias, and completeness of information from the federal government.

Table 3: Trust in information coming from the federal government.

<table>
<thead>
<tr>
<th>Information coming from the federal government (like the president or vice president)</th>
<th>Can be trusted</th>
<th>In accurate</th>
<th>Tells the whole story</th>
<th>Is biased or one sided</th>
</tr>
</thead>
<tbody>
<tr>
<td>Race</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White, non-Hispanic</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Working</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lost Job</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cancer</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diabetes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(Continued)
### Table 3: Trust in information coming from the federal government (Continued)

**Information coming from the federal government (like the president or vice president).**

<table>
<thead>
<tr>
<th>Mental Illness</th>
<th>Can be trusted</th>
<th>Is accurate</th>
<th>Tells the whole story</th>
<th>Is biased or one sided</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>Prob True: 41</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Prob False: 10</td>
<td></td>
<td>0.12 (7.06%)</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>Prob True: 38</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Prob False: 14</td>
<td></td>
<td>0.06 (6.00%)</td>
<td></td>
</tr>
</tbody>
</table>

**Breathing Disorder**

| Yes            | Prob True: 16 |
|                | Prob False: 12 | 0.32 (4.00%)| 0.32 (78.00%)          | 0.32 (18.00%)          |  
| No             | Prob True: 21  |
|                | Prob False: 13 |             | 0.13 (4.00%)           |                        |  

**Other Chronic Condition**

| Yes            | Prob True: 17  |
|                | Prob False: 15 | 0.005**    | 17 (20.00%)            | 17 (20.00%)            |  
| No             | Prob True: 20  |
|                | Prob False: 17 |             | 30 (47.00%)            |                        |  

This pattern held true for information coming from scientists, reflecting an overall level of uncertainty about information from both the federal government and scientists among minority groups compared with White/non-Hispanics. Overall, both groups rated the credibility of information coming from the federal government lower compared with information from scientists. Only 14.97% of White/non-Hispanics and 16.67% of minority participants responded that information from the federal government can be trusted, while 93% of White/non-Hispanics and 77% of minority respondents believed that information from scientists can be trusted. Respondents reporting chronic health conditions, particularly diabetes, were significantly more likely to rate information from the federal government as likely to be false, inaccurate, and not telling the whole story compared with respondents not reporting diabetes as a condition. Males were more likely than females to trust information coming from the federal government, and to rate this information as likely to be true. Respondents who reported losing a job due to COVID-19 were less trusting of information coming from the federal government and less likely to believe this information was accurate than respondents not reporting a job loss. Those who lost a job also rated information coming from scientists as less trustworthy and accurate compared with respondents who did not lose a job.

### Table 4: Trust in information coming from scientists.

**Information coming from scientists.**

<table>
<thead>
<tr>
<th>Race</th>
<th>Can be trusted</th>
<th>Is accurate</th>
<th>Tells the whole story</th>
<th>Is biased or one sided</th>
</tr>
</thead>
<tbody>
<tr>
<td>White, non-Hispanic</td>
<td>169 (93.00%)</td>
<td>159 (88.00%)</td>
<td>125 (69.00%)</td>
<td>29 (16.00%)</td>
</tr>
<tr>
<td>Other</td>
<td>68 (70.00%)</td>
<td>62 (72.00%)</td>
<td>43 (49.00%)</td>
<td>19 (22.00%)</td>
</tr>
</tbody>
</table>

**Age**

| 18-45 | 43 (71.8%) | 41 (74.55%)| 29 (52.73%)           | 14 (25.93%)            |  
| 46-64 | 91 (87.59%)| 82 (76.81%)| 57 (56.11%)           | 15 (24.93%)            |  
| 65+   | 101 (91.53%)| 102 (97.93%)| 84 (71.95%)           | 19 (28.63%)            |  

**Gender**

| Male       | 44 (62.72%)| 42 (84.00%)| 38 (74.51%)           | 15 (29.41%)            |  
| Female     | 19 (88.34%)| 18 (82.73%)| 13 (58.82%)           | 13 (15.14%)            |  

**Working**

| Yes | 124 (80.72%)| 114 (83.21%)| 86 (62.77%)           | 27 (20.00%)            |  
| No  | 116 (86.57%)| 109 (82.58%)| 81 (60.45%)           | 20 (15.04%)            |  

**Lost Job**

| Yes | 32 (74.42%)| 28 (65.12%)| 19 (44.19%)           | 4 (9.52%)              |  
| No  | 137 (90.73%)| 127 (85.23%)| 99 (66.00%)           | 32 (21.62%)            |  

**Cancer**

| Yes | 77 (89.00%)| 77 (90.00%)| 65 (75.00%)           | 15 (17.00%)            |  
| No  | 165 (87.00%)| 148 (79.00%)| 104 (55.00%)           | 33 (18.00%)            |  

665
This work was supported by the National Institutes of Health (NIH), National Library of Medicine G08 LM012689-01; an American Cancer Society (ACS) Research Scholar Grant RS G-17-103-01.
References

14. Rubin GJ, Bakhshi S, Amlot R, Fear N, Potts HWW, Michie S. Health Services and Delivery Research. The design of a survey questionnaire to measure perceptions and behaviour during an influenza pandemic: the Flu TEnphone Survey Template (FluTEST). Southampton (UK): NIHR Journals Library Copyright © Queen’s Printer and Controller of HMSO 2014. This work was produced by Rubin et al. under the terms of a commissioning contract issued by the Secretary of State for Health. This issue may be freely reproduced for the purposes of private research and study and extracts (or indeed, the full report) may be included in professional journals provided that suitable acknowledgement is made and the reproduction is not associated with any form of advertising. Applications for commercial reproduction should be addressed to: NIHR Journals Library, National Institute for Health Research, Evaluation, Trials and Studies Coordinating Centre, Alpha House, University of Southampton Science Park, Southampton SO16 7NS, UK.; 2014.
Automatic Assignment of Radiology Examination Protocols Using Pre-trained Language Models with Knowledge Distillation

Wilson Lau\textsuperscript{1}, Laura Aaltonen, MD PhD\textsuperscript{2}, Martin Gunn, MB ChB\textsuperscript{2}, Meliha Yetisgen, PhD\textsuperscript{1,3}
\textsuperscript{1}Department of Biomedical and Health Informatics, \textsuperscript{2}Department of Radiology, \textsuperscript{3}Department of Linguistics, University of Washington, Seattle, WA

Abstract

Selecting radiology examination protocol is a repetitive, and time-consuming process. In this paper, we present a deep learning approach to automatically assign protocols to computed tomography examinations, by pre-training a domain-specific BERT model (BERT\textsubscript{md}). To handle the high data imbalance across exam protocols, we used a knowledge distillation approach that up-sampled the minority classes through data augmentation. We compared classification performance of the described approach with n-gram models using Support Vector Machine (SVM), Gradient Boosting Machine (GBM), and Random Forest (RF) classifiers, as well as the BERT\textsubscript{base} model. SVM, GBM and RF achieved macro-averaged F1 scores of 0.45, 0.45, and 0.6 while BERT\textsubscript{base} and BERT\textsubscript{md} achieved 0.61 and 0.63. Knowledge distillation boosted performance on the minority classes and achieved an F1 score of 0.66.

Introduction

When an advanced imaging order is placed, e.g. for computed tomography (CT) or magnetic resonance imaging (MRI), a radiologist often has to select the most applicable imaging protocol for the imaging technologist to perform the examination. The imaging protocol contains the technical parameters for CT or MRI image acquisition. For CTs and MRIs, this includes the number of sequences, use of intravenous or oral contrast, scanning planes and other technical parameters to ensure that the image acquisition is best suited to answer the clinical question being asked by the ordering provider. The radiologist gives the protocol decision based on the free-text clinical information in the electronic order and the suggested examination by the ordering physician. This manual protocoling process can be time-consuming, may delay performing timely imaging, and result in unnecessary variability in the techniques used for image acquisition\textsuperscript{1}.

Radiologists are one of the highest paid medical specialties, with an average salary of USD $485,000 in 2020, according to the 2020 physician compensation report published by Doximity. Radiologists’ primary clinical and revenue generating task is image interpretation, but their time is often spent performing ‘non-interpretive tasks’, which are non-revenue generating and often can distract them from performing quality image interpretations. Radiologists may spend 37-44\% of their typical workday performing non-interpretive tasks, and 3.5-6.2\% of their time protocoling\textsuperscript{2,3}. For clinically urgent studies, radiologists may be interrupted to protocol non-urgent examinations, which reduces productivity, potentially impairs interpretive accuracy and lengthens report turnaround time (a radiology service quality metric)\textsuperscript{4,5}.

Radiology exam protocoling is a repetitive and fairly simple classification task. It is therefore a strong candidate for automation. By applying machine learning techniques to protocoling, radiologists could spend a greater proportion of their time performing interpretive tasks, thereby improving the cost-effectiveness of a radiology practice, reducing interruptions for protocoling, improve interpretation accuracy and shorten report turnaround time. Integration of a ML protocoling pipeline into the protocoling software that radiologists use (usually the radiology information system) would permit radiologists to obviate the need to protocol examinations altogether or, in some cases, reduce the number of protocoling steps by suggesting the correct protocol or simply flagging more complex protocols for a radiologist to ‘manually’ protocol.

In this paper, we defined the task of protocol assignment as a classification task. We used structured radiology exam meta-data (exam name and code provided by referring physician) and patient demographics (age and gender) as well as free text diagnoses and history information to automatically assign a radiology protocol. Table 1 presents an example of the radiology examination data from our dataset. Information listed in Table 1 is available to radiologists when they assign protocols manually. In our experiments, we (1) compared different statistical ML models to the state-of-the-art BERT\textsuperscript{6} model for radiology protocol classification task, (2) evaluated the BERT model pre-trained on
general domain ($BERT_{base}$) in comparison to a BERT model pre-trained on our radiology corpus ($BERT_{rad}$), and (3) applied deep learning knowledge distillation approach to tackle high data imbalance in our dataset.

<table>
<thead>
<tr>
<th>Exam metadata</th>
<th>Demographics</th>
<th>Patient history</th>
<th>Protocol</th>
</tr>
</thead>
<tbody>
<tr>
<td>Code</td>
<td>Name</td>
<td>Sex</td>
<td>Age</td>
</tr>
<tr>
<td>CABDWC</td>
<td>CT ABDOMEN W CONTRAST</td>
<td>2</td>
<td>67</td>
</tr>
</tbody>
</table>

Table 1. Example examination data from our dataset.

Related Work

Automating radiology protocol selection has been investigated in previous studies. Brown et al. compared three different ML models, including support vector machine (SVM), gradient boosting machine (GBM), and random forest (RF), to classify MRI protocol selection. They used bag-of-words approach with unigrams to represent features for the text data and combined them with the structured variables (age, sex, location and ordering service). The dataset consisted of 7487 observations. Since each protocol can consist of a sequence of procedures, it is considered a multi-label classification task. They trained 41 binary classifiers for each model to predict each procedure in a sequence. The three ML algorithms included in this study demonstrated similar performance. GBM achieved 86% precision and 80% recall. SVM achieved 83% precision and 82% recall, followed by RF with 85% precision and 80% recall. In another study, the same authors employed similar ML approaches to protocol and prioritize MRI brain examinations. Their best classifier using RF achieved 82% precision and 83% recall. In this paper, we used SVM, GBM, and RF as baselines to compare the performance of our proposed classification approach.

Trivedi et al. used IBM Watson to determine the use of intravenous contrast for musculoskeletal MRI protocols by analyzing only clinical texts. The task was to classify a free-text clinical indication into one of the two labels “with contrast” or “non-contrast”. The dataset consisted of 650 positive and 870 negative labels. Watson achieved over 90% precision and 74% recall. The overall performance is similar to their ensemble model comprising 8 traditional statistical models (SVM, scaled linear discriminant analysis, boosting, bagging, classification and regression tree, RF, Lasso and elastic-net regularized generalized linear model, maximum entropy). Although they claimed that Watson’s classifier was based on deep learning, no specific details about the model architecture and hyperparameters were provided by IBM.

One research conducted by Kalra et al. is the most similar to our study. They developed two statistical ML models and one deep learning model to automate CT and MRI protocol assignment. The dataset contained 18000 CT and MRI examinations in 108 unique protocols. Similar to our dataset, their protocol frequency distribution is highly imbalanced with the 5 most commonly assigned protocols making up 49% of the entire dataset. They trained a k-nearest neighbor and a random forest classifier using TF-IDF feature vectors on unigrams from clinical texts. Interestingly, they excluded structured data elements such as age and gender, which could be strong predictor variables. The performance results from the top two classifiers, RF (80% precision, 82% recall) and DNN (82% precision, 84% recall), were comparable. However, they only reported weighted micro-averages and did not report performance metrics per protocol. Hence, we do not know how the model performed on the minority classes.

Our main contribution in this paper is the feasibility analysis of applying transfer learning using pre-trained language models for protocol classification task. In our experiments, we first evaluated the three ML approaches (SVM, GBM, and RF) presented in previous studies, and observed similar micro-averaged classification performance. Compared to other published datasets, our dataset is relatively larger in which 57% of the data fall into two specific protocol groups. To handle such high data imbalance, we presented a knowledge distillation approach by up sampling the minority classes through data augmentation. We presented the classification result for each protocol group and showed the performance gain for the minority classes.

Methods

Dataset:

Our dataset included 35085 radiology body CT examinations performed at 7 hospital-based and clinic-based imaging sites between January 2018 and June 2019. The data was extracted from the University of Washington radiology
information system. As demonstrated in Table 1, each exam is represented with 4 structured data fields including exam meta-data (exam code, protocol code) and patient demographics (age, gender) as well as 2 unstructured fields to capture patient history (history, diagnosis). Table 2 describes the word level statistics on the two unstructured fields. In our initial analysis of the dataset, we observed that the lengths of the unstructured data are relatively short (average numbers of words for history and diagnosis fields were 8 and 10 with standard deviations 6.57 and 8.6 respectively). 4759 (13.6%) examinations contained no history data and 3 (0.01%) examinations contained no diagnosis data.

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>Max</th>
<th>Mean</th>
<th>Median</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>History</td>
<td>0</td>
<td>47</td>
<td>8</td>
<td>6</td>
<td>6.57</td>
</tr>
<tr>
<td>Diagnosis</td>
<td>0</td>
<td>108</td>
<td>10</td>
<td>8</td>
<td>8.6</td>
</tr>
</tbody>
</table>

**Table 2.** Word statistics on unstructured fields.

In addition, we observed that some of the same protocols had multiple protocol codes, reflecting different codes for the identical protocols performed at different imaging sites. To remove this inconsistency, we manually categorized the protocol codes into 27 unique “protocol groups”; each group unified identical protocols with different codes. We excluded 2 groups that had less than 20 examinations in our experiments (CT CA Oral Only and CT Abdomen IV Only). Table 3 shows the examination frequency with percentages for each protocol group. As can be observed from Table 3, the dataset is highly imbalanced, with the first two protocol groups constituting 57% of the entire dataset. The distribution of examination frequency among the groups has a mean of 1299, median of 200 and standard deviation of 2706.

<table>
<thead>
<tr>
<th>Protocol group</th>
<th>Fre.</th>
<th>%</th>
<th>Protocol group</th>
<th>Fre.</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. CT CAP IV and Oral</td>
<td>11911</td>
<td>33.95%</td>
<td>15. CT Pancreas Mass 3 Phase</td>
<td>202</td>
<td>0.58%</td>
</tr>
<tr>
<td>2. CT Abdomen Pelvis w IV Only</td>
<td>8057</td>
<td>22.96%</td>
<td>16. CT Abdomen No Contrast</td>
<td>195</td>
<td>0.56%</td>
</tr>
<tr>
<td>3. CT CAP IV Only</td>
<td>3351</td>
<td>9.55%</td>
<td>17. CT CA IV and Oral</td>
<td>194</td>
<td>0.55%</td>
</tr>
<tr>
<td>4. CT Abdomen Pelvis w IV and Oral</td>
<td>2941</td>
<td>8.38%</td>
<td>18. CT Pelvis IV Only</td>
<td>192</td>
<td>0.55%</td>
</tr>
<tr>
<td>5. CT Renal Mass</td>
<td>2036</td>
<td>5.80%</td>
<td>19. CT Abdomen IV and Oral</td>
<td>173</td>
<td>0.49%</td>
</tr>
<tr>
<td>6. CT Liver 3 Phase</td>
<td>1652</td>
<td>4.71%</td>
<td>20. CT Pancreas Mass 2 Phase</td>
<td>143</td>
<td>0.41%</td>
</tr>
<tr>
<td>7. CT Abdomen Pelvis No Contrast</td>
<td>931</td>
<td>2.65%</td>
<td>21. CT Abdomen Pelvis w Oral only</td>
<td>132</td>
<td>0.38%</td>
</tr>
<tr>
<td>8. CT IVP 50 yrs +</td>
<td>854</td>
<td>2.43%</td>
<td>22. CT CA No Contrast</td>
<td>75</td>
<td>0.21%</td>
</tr>
<tr>
<td>9. CT CAP Oral Only</td>
<td>531</td>
<td>1.51%</td>
<td>23. CT Pelvis Cystogram</td>
<td>68</td>
<td>0.19%</td>
</tr>
<tr>
<td>10. CT CAP No Contrast</td>
<td>336</td>
<td>0.96%</td>
<td>24. CT Liver 2 Phase</td>
<td>51</td>
<td>0.15%</td>
</tr>
<tr>
<td>11. CT Abd Pel Enterography</td>
<td>297</td>
<td>0.85%</td>
<td>25. CT Pelvis IV and Oral</td>
<td>42</td>
<td>0.12%</td>
</tr>
<tr>
<td>12. CT Liver 4 Phase</td>
<td>252</td>
<td>0.72%</td>
<td>26. CT CA Oral Only (excluded)</td>
<td>15</td>
<td>0.04%</td>
</tr>
<tr>
<td>13. CT CA IV Only</td>
<td>226</td>
<td>0.64%</td>
<td>27. CT Abdomen IV Only (excluded)</td>
<td>8</td>
<td>0.02%</td>
</tr>
<tr>
<td>14. CT IVP &lt; 50</td>
<td>220</td>
<td>0.63%</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 3.** Distribution of examinations across protocols.

**Approach:**

We trained a deep learning classifier using the state-of-the-art neural language model, BERT® to automatically assign protocols to computer tomography (CT) examinations. Specifically, we fine-tuned the Google pre-trained model BERTbase with a linear layer on top using cross-entropy loss. We formulated the task as a single-sequence classification task by first transforming the structured and unstructured data into the following template: “Exam is <exam code>. Sex is <gender>. Age at Exam <age>. History: <history>. Diagnosis: <diagnosis>” and subsequently classifying it into one of 25 protocol groups listed in Table 3. We observed that the mean and median of number of characters in the templated data are 192 and 178. In order to capture context presented in the training instances, we set the maximum
sequence length parameter of the BERT model to be 200 with a batch size of 48. We followed the suggestions described in the BERT paper and used the Adam optimizer with a learning rate of 2-e5. We fine-tuned the BERT model for 4 epochs.

Conceptually, BERT learns the relations between words by randomly masking words in a sequence with a [MASK] token and then trains itself to predict them from the context of the unmasked ones. Additionally, it learns the sentence relationships by training itself to predict if the second sentence in a pair is truly following the first sentence in the corpus. These two learning tasks allow BERT to self-train and capture the context of language used in an unlabeled corpus before transferring all parameters to downstream applications. Previous studies showed promising results of using BERT in clinical applications. Examples include chest x-ray reports classification\textsuperscript{11}, and relation extraction in clinical and biomedical domain\textsuperscript{12,13}. Since BERT\textsubscript{base} was originally pre-trained on BookCorpus and English Wikipedia, to fully encode the semantic context in clinical and biomedical text, it has been shown that further training BERT\textsubscript{base} on MIMIC and PubMed data can boost the performance of named entity recognition in the biomedical domain\textsuperscript{14,15}. Inspired by these studies, we further pre-trained BERT\textsubscript{base} on our radiology protocol corpus and named it BERT\textsubscript{rad}. We repeated the same experiment with BERT\textsubscript{rad} using the same hyperparameters listed above. All BERT experiments were implemented with Huggingface’s transformer library\textsuperscript{16}.

Knowledge distillation:

Imbalanced class distribution usually leads to poor classification results on the minority classes\textsuperscript{17}. When dealing with imbalanced datasets, a popular approach is to use the Synthetic Minority Oversample Technique (SMOTE)\textsuperscript{18} which generates new artificial samples for the minority classes by interpolating the nearest neighbors of the existing samples. This method reduces the likelihood of overfitting minority classes commonly observed in random over sampling approach. However, because the inputs of the BERT model include positional embeddings and WordPiece tokenization with special classifier token [CLS] and separator token [SEP], synthesizing these input values in vector space using interpolation will lose the context of the tokens in the samples.

Recent studies have successfully demonstrated the possibility to transfer task specific knowledge from the large BERT model to a smaller neural architecture without significant performance degradation\textsuperscript{19-21}, using a technique called knowledge distillation. The process involves training a second model (student) to match the predictions from the first model (teacher). We hypothesized that by transferring knowledge specific to the minority classes from the BERT\textsubscript{rad} model to a second BERT model, we could improve the classification performance on the minority classes. In particular, we aimed to train a student model that could outperform the teacher with identical neural architecture. Furlanello et al. referred to this approach as Born-Again Neural Network (BAN)\textsuperscript{22}, which has been shown to produce better results in both single and multi-task settings\textsuperscript{23}. During the knowledge distillation process, the raw predictions from the teacher model, known as logits, are being used as “soft labels” for training the student model. As Hinton et al. suggest, the distribution in the logits, even among incorrect predictions, contains information about how the teacher model is generalizing, thereby offering more training signals than one-hot categorical labels\textsuperscript{19}.

To effectively transfer knowledge about the minority classes to a student model, a large unlabeled dataset is needed to generate enough soft labels from the teacher model. In this study, we applied Tang et al.’s data augmentation techniques to synthesize masked data in order to allow the teacher to fully express its knowledge\textsuperscript{24}. To augment a given training instance, we randomly sampled a number $P$ from the uniform distribution $[0,1]$. If $P < 0.1$, we randomly replaced a word in the history and diagnosis section with the [MASK] token. If $P$ is between 0.1 and 0.2, we randomly replaced a word with another word in the training set that has the same POS tag. Finally, we randomly replaced an n-gram ($n \in [1,3]$) in the training instance with the [MASK] token. This technique is similar to the masking procedure employed in BERT’s masked language model. We repeated this augmentation process to generate 30 new instances, without duplication, for each training instance. We evaluated different numbers of augmented instances (25, 30, 35, 40, 50) by running 5-fold cross validation with the augmented data. Our evaluation showed that the experiment with 30 augmented instances achieved the best result. In addition, we wanted to limit the augmented sample size of the dominant classes and therefore set a maximum sampling limit of 12000, such that the final sample size of each class after augmentation could not exceed 12000. We then ran inferencing on the augmented dataset using the teacher model BERT\textsubscript{rad} to generate soft labels for distillation. Finally, we initialized a student BERT\textsubscript{rad} model with a different random seed and trained it to imitate the teacher by minimizing the mean squared error (MSE) between the student’s logits and teacher’s logits. At the same time, we allowed the student model to surpass the teacher by training with the true labels by minimizing the cross-entropy loss against the one-hot multi-class labels:

$$L_{distill} = \alpha \cdot L_{cross-entropy} + (1 - \alpha) \cdot L_{MSE}$$
where \( \alpha \) is the ratio of true labels within a single batch of training samples. After each iteration of knowledge distillation, the student model became the teacher for next generation.

Baselines: For our prediction task, we trained three separate ML models with Support Vector Machine (SVM), Gradient Boosting Machine (GBM) and Random Forest (RF) as baselines and compared their performance against our proposed deep learning classifier. To train the baselines, we transformed the unigrams and bigrams of the history and diagnosis notes into feature space using TF-IDF before combining with the numeric values in the structured data. The baselines were trained with the same features. All statistical classification modelling was implemented using the Scikit-learn machine learning python package\(^2\).

Results

We used 5-fold cross validation to evaluate the general performance of the models. For each fold, the models were trained on the same training data and evaluated on the same held-out test data. We used precision, recall, and F1-score as metrics to measure the performance. The overall macro-average and weighted micro-average results are presented in Table 4. To generate macro-averaged results, we first calculated the metrics for each class and calculated the average of them giving each class equal weight. To generate weighted micro-averaged metrics, we calculated the metric averages weighted by the number of true labels in each class. As can be observed, the micro-average results are largely similar due to the bias towards the majority classes. In the macro-average results, among the baselines, RF performed the best with 0.60 F1-Score. Both SVM and GBM produced 0.45 F1-score. The SVM in general produced higher precision and lower recall, when compared to GBM. The classifiers based on BERT models performed better than the SVM, GBM and RF baselines. Furthermore, the in-domain BERT\(_{rad}\) produced 0.2 higher macro F1 score than the out-of-domain BERT\(_{base}\) model (0.63 versus 0.61).

To mitigate the high data imbalance, two resampling experiments were performed with the BERT\(_{rad}\) model. The resampling was performed only on the training data while the validation data were kept the same. First, we undersampled the 2 majority classes by randomly removing some training instances such that their sample sizes matched the size of the third largest protocol group (#3). The result shows that the macro-average F1 dropped 0.24 and the weighted micro-average F1 dropped 0.2 due to the misclassification of the majority classes given their smaller sample sizes. In the over-sampling experiment, the training instances in the minority classes were randomly replicated so that their sample sizes matched the size of the second largest protocol group (#2). The result shows no performance improvement in the macro-average F1 but degradation in the weighted micro-average. This can be caused by overfitting the large number of duplicate training samples in the minority classes\(^18\). The results also show that the BAN models \{2,3\} achieved better macro-average performance than BERT\(_{base}\) and BERT\(_{rad}\), without any degradation in micro-average performance. More specifically, the macro-average F1 in generations of student models (BAN\{1,2,3\}) improved, suggesting that the classifiers achieved better performance in predicting the minority classes through knowledge distillation. We also observed that the performance saturated after training the second generation of BAN student model. This finding is similar to the one reported by Furlanello et al\(^22\).

<table>
<thead>
<tr>
<th>Model</th>
<th>Macro average</th>
<th>Micro (Weighted) average</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td>SVM</td>
<td>0.60</td>
<td>0.42</td>
</tr>
<tr>
<td>GBM</td>
<td>0.46</td>
<td>0.46</td>
</tr>
<tr>
<td>RF</td>
<td>0.63</td>
<td>0.59</td>
</tr>
<tr>
<td>BERT(_{base})</td>
<td>0.68</td>
<td>0.60</td>
</tr>
<tr>
<td>BERT(_{rad})</td>
<td>0.67</td>
<td>0.62</td>
</tr>
<tr>
<td>BERT(_{rad}) undersample</td>
<td>0.42</td>
<td>0.38</td>
</tr>
<tr>
<td>BERT(_{rad}) oversample</td>
<td>0.63</td>
<td>0.63</td>
</tr>
<tr>
<td>BAN1</td>
<td>0.68</td>
<td>0.64</td>
</tr>
<tr>
<td>BAN2</td>
<td>0.69</td>
<td>0.65</td>
</tr>
<tr>
<td>BAN3</td>
<td>0.69</td>
<td>0.65</td>
</tr>
</tbody>
</table>

Table 4. Comparison of model results. BAN\{1,2,3\} denotes the 1st, 2nd and 3rd generation of knowledge distillation.
Table 6 presents performance results at the protocol level. Our results showed that the BERT models generally outperformed statistical baselines among the protocol groups. From the protocol group #2 to #9, the difference in F1 scores between the best performing statistical model and the best performing BERT model is between 1% to 5%. In some specific groups, such as “CT CAP No Contrast” (#10) and “CT Abd Pel Enterography” (#11), we observed much larger improvement. Overall, we did not observe a substantial improvement in BERT_rad compared to BERT_base. The small size of the dataset and sparseness of the unstructured data fields resulted this outcome. Nonetheless, BERT_rad was able to outperform BERT_base in some protocol groups by capturing the context of words that are not common in general domain. For example, in the protocol group “CT Abd Pel Enterography” (#11), the word hernia, which describes the symptom that a tissue pushes through the abdominal opening, appeared in over 79% of the diagnosis fields, while another word CREATININE, a compound that indicates the level of kidney function, appeared in over 73% of the history fields. These two medical terms are not commonly seen in the general corpora. By pre-training on the radiology corpus, BERT_rad was able to learn better contextual representation of these medical terms and outperformed BERT_base by 0.07 F1 in that protocol group. We observed similar improvement in protocol groups “CT Abdomen IV and Oral” (#19) and “CT Abdomen Pelvis w Oral only” (#21).

One interesting observation was BERT_base model’s substantially low F1-score of 0.16 for group “CT IVP < 50” (#14) when compared to the F1-scores (SVM: 0.61, GBM: 0.73, RF: 0.88) of statistical baselines. Further investigation showed that 87% of the false negatives for “CT IVP < 50” (#14) were misclassified to “CT IVP 50 yrs +” (#8) by BERT_base. Because the main difference between these two protocol groups is age of patient, the age feature by itself offered high information gain to allow RF to learn a more robust model. On the other hand, deep learning models require considerably large volume of data to extract patterns from high-dimensional data points. The smaller sample size of protocol group #14 limited the BERT_base model to learn to differentiate from protocol group #8. However, data augmentation in the knowledge distillation process eventually supplied additional training signals for the model to generalize, leading to the similar performance levels as RF.

Although knowledge distillation enabled the BERT models to improve overall performance on the minority classes, one particular protocol group that was not correctly classified by any models was “CT Liver 2 Phase” (#24). Our error analysis showed that the models misclassified some #24 cases to “CT Liver 3 Phase” (#6) because of similar patient diagnosis and history. Table 5 presents one of these cases.

<table>
<thead>
<tr>
<th>Protocol group</th>
<th>History</th>
<th>Diagnosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>6. CT Liver 3 Phase</td>
<td>Last creatine level:CREATININE 0.92</td>
<td>ABDOMEN W/CONTRAST; 6MO REPEAT F/U FOR HCC SURVEILLANCE, S/P LIVER TRANSPLANT</td>
</tr>
<tr>
<td>24. CT Liver 2 Phase</td>
<td>Last creatine level:CREATININE 0.81</td>
<td>ABDOMEN W/CONTRAST; TO EVALUATE SIZE OF PSEUDOCYST, S/P LIVER TRANSPLANT</td>
</tr>
</tbody>
</table>

Table 5. Examinations in two different protocol groups with similar history and diagnosis.

While these were the correct protocol assignments in clinical practice, because #24 only constituted 0.15% of the training data and was 30 times less than #6, there were not enough data to train the models to differentiate #24 from #6. Additionally, we found that some #24 cases were misclassified to “CT CAP IV and Oral” (#1) because of the exact same history and diagnosis found in #24. For instance, there were 6 cases with history of “ORAL and IV Contrast” and diagnosis of “2 Phase Liver, PNET Metastatic” assigned to protocol #1 and 1 case with the same history and diagnosis assigned to #24. Without any additional clinical information to help differentiate the two protocol assignments, the models simply inferred to the group that was more dominant in the training data.
source codes will be shared with radiologist workflow and used on biomedical datasets are intrinsically imbalanced (e.g. the prevalence of certain diseases). Improved overall classification performance of models assigned protocols based on patient demographic and history data. The results showed that overall pre-trained language models performed better than traditional n-gram models. Additionally, we demonstrated that knowledge distillation improved overall classification performance for the majority of under-represented groups. Since many real-world biomedical datasets are intrinsically imbalanced (e.g. the prevalence of certain uncommon cancer types or chronic diseases), we think that this technique could be useful in many classification problems involving clinical text using pre-trained language models. In future studies, we plan to integrate the machine learning models into actual radiologist workflow and compare the agreement between radiologists and models. The dataset and experimentation source codes will be shared with the research community.

Table 6. Comparison of model results in F1 for each protocol group on the test set.

<table>
<thead>
<tr>
<th>Protocol group</th>
<th>Exam count</th>
<th>SVM</th>
<th>GBM</th>
<th>RF</th>
<th>BERT\textsubscript{base}</th>
<th>BERT\textsubscript{rad}</th>
<th>BAN1</th>
<th>BAN2</th>
<th>BAN3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. CT CAP IV and Oral</td>
<td>2382</td>
<td>0.92</td>
<td>0.91</td>
<td>0.93</td>
<td>0.93</td>
<td>0.93</td>
<td>0.93</td>
<td>0.93</td>
<td></td>
</tr>
<tr>
<td>2. CT Abdomen Pelvis w IV Only</td>
<td>1612</td>
<td>0.85</td>
<td>0.86</td>
<td>0.87</td>
<td>0.88</td>
<td>0.87</td>
<td>0.87</td>
<td>0.87</td>
<td></td>
</tr>
<tr>
<td>3. CT CAP IV Only</td>
<td>670</td>
<td>0.78</td>
<td>0.73</td>
<td>0.78</td>
<td>0.80</td>
<td>0.80</td>
<td>0.79</td>
<td>0.79</td>
<td>0.80</td>
</tr>
<tr>
<td>4. CT Abdomen Pelvis w IV and Oral</td>
<td>588</td>
<td>0.66</td>
<td>0.66</td>
<td>0.66</td>
<td>0.70</td>
<td>0.69</td>
<td>0.67</td>
<td>0.67</td>
<td>0.67</td>
</tr>
<tr>
<td>5. CT Renal Mass</td>
<td>407</td>
<td>0.83</td>
<td>0.89</td>
<td>0.91</td>
<td>0.92</td>
<td>0.93</td>
<td>0.92</td>
<td>0.93</td>
<td>0.92</td>
</tr>
<tr>
<td>6. CT Liver 3 Phase</td>
<td>330</td>
<td>0.83</td>
<td>0.82</td>
<td>0.85</td>
<td>0.87</td>
<td>0.87</td>
<td>0.86</td>
<td>0.86</td>
<td>0.87</td>
</tr>
<tr>
<td>7. CT Abdomen Pelvis No Contrast</td>
<td>186</td>
<td>0.42</td>
<td>0.73</td>
<td>0.75</td>
<td>0.77</td>
<td>0.76</td>
<td>0.77</td>
<td>0.77</td>
<td>0.77</td>
</tr>
<tr>
<td>8. CT IVP 50 yrs +</td>
<td>171</td>
<td>0.81</td>
<td>0.90</td>
<td>0.92</td>
<td>0.84</td>
<td>0.84</td>
<td>0.91</td>
<td>0.93</td>
<td>0.92</td>
</tr>
<tr>
<td>9. CT CAP Oral Only</td>
<td>107</td>
<td>0.29</td>
<td>0.56</td>
<td>0.31</td>
<td>0.58</td>
<td>0.59</td>
<td>0.59</td>
<td>0.59</td>
<td>0.61</td>
</tr>
<tr>
<td>10. CT CAP No Contrast</td>
<td>67</td>
<td>0.01</td>
<td>0.01</td>
<td>0.07</td>
<td>0.16</td>
<td>0.16</td>
<td>0.14</td>
<td>0.13</td>
<td>0.16</td>
</tr>
<tr>
<td>11. CT Abd Pel Enterography</td>
<td>59</td>
<td>0.54</td>
<td>0.41</td>
<td>0.50</td>
<td>0.53</td>
<td>0.60</td>
<td>0.63</td>
<td>0.61</td>
<td>0.62</td>
</tr>
<tr>
<td>12. CT Liver 4 Phase</td>
<td>51</td>
<td>0.07</td>
<td>0.33</td>
<td>0.58</td>
<td>0.66</td>
<td>0.69</td>
<td>0.68</td>
<td>0.67</td>
<td>0.67</td>
</tr>
<tr>
<td>13. CT CA IV Only</td>
<td>45</td>
<td>0.59</td>
<td>0.49</td>
<td>0.75</td>
<td>0.78</td>
<td>0.78</td>
<td>0.77</td>
<td>0.78</td>
<td>0.78</td>
</tr>
<tr>
<td>14. CT IVP &lt; 50</td>
<td>44</td>
<td>0.61</td>
<td>0.73</td>
<td>0.88</td>
<td>0.88</td>
<td>0.88</td>
<td>0.88</td>
<td>0.88</td>
<td>0.87</td>
</tr>
<tr>
<td>15. CT Pancreas Mass 3 Phase</td>
<td>41</td>
<td>0.46</td>
<td>0.36</td>
<td>0.58</td>
<td>0.64</td>
<td>0.62</td>
<td>0.63</td>
<td>0.67</td>
<td>0.65</td>
</tr>
<tr>
<td>16. CT Abdomen No Contrast</td>
<td>39</td>
<td>0.61</td>
<td>0.43</td>
<td>0.79</td>
<td>0.82</td>
<td>0.81</td>
<td>0.81</td>
<td>0.82</td>
<td>0.80</td>
</tr>
<tr>
<td>17. CT CA IV and Oral</td>
<td>39</td>
<td>0.12</td>
<td>0.10</td>
<td>0.56</td>
<td>0.61</td>
<td>0.62</td>
<td>0.61</td>
<td>0.61</td>
<td>0.61</td>
</tr>
<tr>
<td>18. CT Pelvis IV Only</td>
<td>39</td>
<td>0.78</td>
<td>0.77</td>
<td>0.80</td>
<td>0.83</td>
<td>0.83</td>
<td>0.82</td>
<td>0.83</td>
<td>0.83</td>
</tr>
<tr>
<td>19. CT Abdomen IV and Oral</td>
<td>35</td>
<td>0.06</td>
<td>0.03</td>
<td>0.36</td>
<td>0.41</td>
<td>0.45</td>
<td>0.48</td>
<td>0.48</td>
<td>0.48</td>
</tr>
<tr>
<td>20. CT Pancreas Mass 2 Phase</td>
<td>28</td>
<td>0.05</td>
<td>0.15</td>
<td>0.21</td>
<td>0.29</td>
<td>0.28</td>
<td>0.28</td>
<td>0.29</td>
<td>0.29</td>
</tr>
<tr>
<td>21. CT Abdomen Pelvis w Oral only</td>
<td>26</td>
<td>0.00</td>
<td>0.05</td>
<td>0.26</td>
<td>0.23</td>
<td>0.37</td>
<td>0.39</td>
<td>0.38</td>
<td>0.37</td>
</tr>
<tr>
<td>22. CT CA No Contrast</td>
<td>15</td>
<td>0.19</td>
<td>0.09</td>
<td>0.71</td>
<td>0.76</td>
<td>0.75</td>
<td>0.74</td>
<td>0.75</td>
<td>0.76</td>
</tr>
<tr>
<td>23. CT Pelvis Cystogram</td>
<td>14</td>
<td>0.69</td>
<td>0.27</td>
<td>0.95</td>
<td>0.95</td>
<td>0.95</td>
<td>0.95</td>
<td>0.95</td>
<td>0.96</td>
</tr>
<tr>
<td>24. CT Liver 2 Phase</td>
<td>10</td>
<td>0.00</td>
<td>0.04</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>25. CT Pelvis IV and Oral</td>
<td>8</td>
<td>0.00</td>
<td>0.03</td>
<td>0.03</td>
<td>0.15</td>
<td>0.31</td>
<td>0.28</td>
<td>0.25</td>
<td>0.28</td>
</tr>
</tbody>
</table>

Conclusion

In this study, we presented a novel ML approach using pre-trained language models to help radiologists automatically assign protocols based on patient demographic and history data. The results showed that overall pre-trained language models performed better than traditional n-gram models. Additionally, we demonstrated that knowledge distillation improved overall classification performance for the majority of under-represented groups. Since many real-world biomedical datasets are intrinsically imbalanced (e.g. the prevalence of certain uncommon cancer types or chronic diseases), we think that this technique could be useful in many classification problems involving clinical text using pre-trained language models. In future studies, we plan to integrate the machine learning models into actual radiologist workflow and compare the agreement between radiologists and models. The dataset and experimentation source codes will be shared with the research community.

* https://github.com/wilsonlau-wu/Automatic-Assignment-of-Radiology-Examination
Acknowledgements

This publication was partially supported by the National Center For Advancing Translational Sciences of the National Institutes of Health under Award Number UL1 TR002319. Research and results reported in this publication was facilitated by the generous contribution of computational resources from the University of Washington Department of Radiology.

References

Hybrid Ensemble-Rule Algorithm for Improved MEDLINE® Sentence Boundary Detection

Daniel X. Le, James G. Mork, and Sameer Antani
Lister Hill National Center for Biomedical Communications
National Library of Medicine
8600 Rockville Pike, Bethesda, MD 20894

Abstract

Sentence boundary detection (SBD) is a fundamental building block in the Natural Language Processing (NLP) pipeline. Incorrect SBD may impact subsequent processing stages resulting in decreased performance. In well-behaved corpora, a few simple rules based on punctuation and capitalization are sufficient for successfully detecting sentence boundaries. However, a corpus like MEDLINE citations presents challenges for SBD due to several syntactic ambiguities, e.g., abbreviation-periods, capital letters in first words of sentences, etc. In this manuscript we present an algorithm to address these challenges based on majority voting among three SBD engines (Python NLTK, pySBD, and Syntok) followed by custom post-processing algorithms that rely on NLP spaCy part-of-speech, abbreviation and capital letter detection, and computing general sentence statistics. Experiments on several thousand MEDLINE citations show that our proposed approach for combining multiple SBD engines and post-processing rules performs better than each individual engine.

Introduction

This paper describes an ongoing effort at the National Library of Medicine (NLM) related to the detection of sentence boundaries of MEDLINE documents where a document refers to a MEDLINE citation which contains a title and may contain an abstract.

Sentence boundary detection (SBD) is an important pre-processing step in many Natural Language Processing (NLP) applications, such as sentence embedding, part-of-speech taggers, document indexing, and question answering. Several techniques for detecting sentence boundaries have been reported in the NLP literature covering a variety of text domains including general text (WSJ text and the Brown corpus [1,2]), biomedical text (the GENIA corpus [3] of biomedical abstracts), clinical text (the i2b2 corpus [4]), and legal text (the United States Courts decisions dataset [5]).

As pointed out in reviews by Griffis [6] and Read [7], the performance of the reviewed SBD techniques is domain dependent. Most do well in general text domains that “conform to formal English” but perform poorly in certain domains, like biomedicine and legal text, where customized algorithms tend to perform better. A specialized corpus like MEDLINE containing journal citations, titles and abstracts presents challenges for SBD due to the difficulties in recognizing uppercase letters belonging to domain-specific (genetic, biological, and chemical) terms, and disambiguating abbreviation-periods among journal abbreviations and medical terms.

In the domain of general text SBD, Riley [8] used a decision tree classifier while Gillick [9] used a supervised approach with Naïve Bayes and Support Vector Machine (SVM). Kiss [10] proposed an unsupervised abbreviation detection algorithm with the Punkt model that “can detect abbreviations based on collocation between periods and truncated words”. Palmer [11] used a neural network with inputs as vectors of binary values representing contextual words surrounding punctuation marks. These techniques reported high accuracies of over 98.00% on both WSJ and the Brown corpus.

Savelka [5] and Sanchez [12] suggested methods to detect sentence boundaries in legal text domains “using customized Punkt tokenizer and Conditional Random Field (CRF) algorithms”. Both systems reported accuracies in the range from 75.00% to 90.00%. Buyko [13] recommended a new approach by retraining the OpenNLP software application on specific biology domains, including the GENIA and PennBioIE25 corpora, and the system reported accuracies of 99.00% and 97.40%, respectively. This experiment was applied on the GENIA corpus [3], a collection with 1,999 abstracts from the MEDLINE database. Kreuzthaler [14] suggested a system using two SVM binary classifiers for abbreviation and sentence detection to detect sentence boundaries in German clinical narratives. The system reported an F-measure of 95.00% for abbreviation detection and 94.00% for sentence delineation.

In this paper, we describe our hybrid-ensemble approach to detect sentence boundaries to address and resolve identified SBD challenges. The algorithm is based on majority voting among three SBD engines (Python NLTK [15], pySBD [16], and Syntok [17]) followed by custom post-processing rule-based algorithms to detect sentence
boundaries that rely on NLP spaCy part-of-speech (POS) [18] tokenization library, abbreviation, and capital letter detection, and computing general sentence statistics. The performance of the proposed SBD algorithm has been evaluated using two data sets of approximately three thousand MEDLINE citations. The evaluation results show that our hybrid-ensemble approach performs better than each individual engine in terms of improving the detection accuracy of correct sentence boundaries.

**Methods**

**Data sets**

The training data set consists of 1,514,446 citations randomly selected from the 30-million citations in the 2020 MEDLINE Baseline data set [19] and is used for observations and for developing features and rules for the proposed SBD algorithm. There are two ground-truth data sets that we used to evaluate the system performance of the proposed SBD algorithm: the GENIA corpus [3] ground-truth data set and the 2021 SBD ground-truth data set.

The GENIA corpus [3] ground-truth data set consists of 18,541 sentences segmented from 1,999 citations from MEDLINE database. It was created from the Genome Informatics Extraction (GENIA) project to provide the gold standard for the evaluation of text mining systems. For the 2021 SBD ground-truth data set, we downloaded approximately 32-million citations from the 2021 MEDLINE Baseline data set [19] in XML file format. We then randomly selected 1,020 citations that were not in the 2020 MEDLINE Baseline data set and extracted titles and abstracts from the XML. For any “structured abstracts” citations, we kept only the label followed by a colon space and filtered out “NlmCategory” and their XML tag “AbstractText” [20]. For example, the XML text “<AbstractText Label="abstracts" citations, we kept only the label followed by a colon space and filtered out

**Basic definitions**

Definitions of the basic features used in our system are given here.

1. A “regular” capital word versus a “special” capital word: In our system, words that begin a sentence could be considered as “regular” or “special” capital words. A “regular” capital word has its first character as an uppercase letter (e.g. “National”) while a “special” capital word includes a lowercase letter after its first character (e.g. “20-Hydroxycholesterol”).

2. **VERB-Phrase sentences and NOUN-Phrase sentences**: POS tagging is a technique to map words as nouns, verbs, adjectives, adverbs, etc. In our system, NLP spaCy POS tagging library [18] is used to generate two lists of attributes of the tokenization outputs for any sentence: a POS list and a TAG list. A POS list shows the coarse-grained part of speech, and its associated TAG list shows the fine-grained part of speech, as shown in the following example.

   Sentence: “This method shows advantages in two aspects.”

A complete list of POS and associated TAG list are available [18]. Based on these lists, a “VERB-Phrase sentence” and a “NOUN-Phrase sentence” are defined as follows:

A sentence is considered a “VERB-Phrase sentence” if it satisfies one of the following two cases:

- Case 1: POS = AUX and TAG = VB or VBD or VBP or VBZ
- Case 2: POS = VERB and TAG = VBD or VBP or VBZ or VBN or MD

A sentence is considered a “NOUN-Phrase sentence” if it satisfies one of the following three cases:

- Case 1: POS = NOUN or PROPN
- Case 2: POS = PRON and TAG = EX or PRP or WP
- Case 3: POS = DET and TAG = WPS

As a result, we define the “five-sentence-based POS tagging types” as follows:

<table>
<thead>
<tr>
<th>Type</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>NP:</td>
<td>A sentence with only a “NOUN-Phrase sentence”, e.g., “Level of evidence : Level II.”</td>
</tr>
<tr>
<td>VP:</td>
<td>A sentence with only a “VERB-Phrase sentence”, e.g., “were investigated.”</td>
</tr>
<tr>
<td>NP+VP:</td>
<td>A sentence with a “NOUN-Phrase sentence” followed by a “VERB-Phrase sentence”, e.g., “Fast and scalable tools are hence needed”</td>
</tr>
<tr>
<td>VP+NP:</td>
<td>A sentence with a “VERB-Phrase sentence” followed by a “NOUN-Phrase sentence”, e.g., “Excluded were 139 Web sites.”</td>
</tr>
<tr>
<td>NP+VP+NP:</td>
<td>A sentence with a “NOUN-Phrase sentence” followed by a “VERB-Phrase” and a “NOUN-Phrase” e.g., “Costs were estimated in US dollars.”</td>
</tr>
</tbody>
</table>
3. Normal POS versus Special POS: Based on our observations, we recognized two POS terms that play an important role in helping to make the connection decisions between consecutive sentences: “PROPN” and “INTJ”. They are defined in this paper as special-POS and the remaining terms are normal-POS (“ADJ”, “ADP”, “ADV”, “AUX”, “CCONJ”, “DET”, “NOUN”, “NUM”, “PART”, “PRON”, “PUNCT”, “SCONJ”, “SYM”, “VERB”, and “X”). The following show examples of broken sentences that could be connected through the help of special-POS.

<table>
<thead>
<tr>
<th>By contrast, cyto.</th>
<th>o, i.e., the ... tension.</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADP, NOUN, PUNCT, PROPN, PUNCT</td>
<td>INTJ, PUNCT, X, PUNCT, DET, ..., NOUN, PUNCT</td>
</tr>
<tr>
<td>Furthermore, ... anion (O2-)</td>
<td>with 1 ... scavenging O2^-</td>
</tr>
<tr>
<td>ADV, PUNCT, ... NOUN, PUNCT, INTJ, PUNCT, PUNCT</td>
<td>ADP, NUM, ..., VERB, X, PUNCT</td>
</tr>
<tr>
<td>PROPN, PROPN, PUNCT</td>
<td>NUM, PUNCT, NUM, PUNCT, PUNCT, NUM, SYM, NUM, PUNCT</td>
</tr>
</tbody>
</table>

### Algorithms to Handle Sentence Boundaries Challenges in MEDLINE

We used two algorithms to address the above-mentioned challenges: the “special” capital word detection algorithm and the abbreviations detection algorithm.

1. The “special” capital word detection algorithm: A sentence usually starts with a word in which its first character is an uppercase letter. However, this condition might not hold true for some genetic, biological, and chemical terms where digits, punctuation, lowercase letters, and symbols (such as “12α-Hydroxyl”, “(-)-Deprenyl”, “•Lichen”) might start a sentence. For these terms, we observed some special patterns from characters before their first uppercase letter (such as “digits, Greek letters, punctuations” for “12α-Hydroxyl”, “punctuations” for “(-)-Deprenyl”, “symbols” for “•Lichen”). Therefore, in order to address such variants, we built a predefined list of “special” capital word patterns and developed an algorithm using this list to detect whether the first word of a sentence is a “special” capital word.

To build a predefined list of capital word patterns, we collected only sentences of which the first word has at least one uppercase letter, but its first character is not an uppercase letter. In addition, we limited the sentence length to be greater or equal to the average sentence length which is 110 characters to reduce the number of collected samples.

Note that a common plain English guideline says that an average sentence length is about 15–20 words [21] and the average word length is about 4.7 characters [22]. In this paper, the average sentence length (AVGSL) is calculated based on 5 characters per word, an average of 18 words, and 17 spaces per sentence as follows:

\[
\text{AVGSL} = 18 \text{ words} \times 5 \text{ characters} + 17 \text{ spaces} \approx 110 \text{ characters.}
\]

To build a predefined list of capital word patterns, we first collected all “special” capital word patterns that are generated by (a) taking the sentences’ first words and truncate them at their uppercase letter. For example, the first word “12α-Hydroxyl” is truncated into “12α”-, (b) building lists of Unicode names associated with characters of the truncated words. For the truncated word “12α-”, the list of Unicode names is: “DIGIT ONE”, “DIGIT TWO”, “GREEK SMALL LETTER ALPHA”, “HYPHEN-MINUS”, (c) replacing Unicode names based on their type by their predefined “replacement” symbols using the replacement table (Table 1). For the list of Unicode names in b), the replacement symbols are 00XH, (d) replacing consecutive similar “replacement” symbols by a single “replacement” symbol to build “special” capital word patterns for the truncated words. For the above replacement symbols in c), the “special” capital word pattern is 0XH. We then tabulated and sorted the collection of the “special” capital word patterns based on their pattern occurrence frequency and selected the topmost patterns as the predefined list of “special” capital word patterns.

### Table 1. Replacement symbols

<table>
<thead>
<tr>
<th>Type of Unicode Name</th>
<th>Replacement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Greek alphabet</td>
<td>X</td>
</tr>
<tr>
<td>Hyphen</td>
<td>H</td>
</tr>
<tr>
<td>Punctuation</td>
<td>P</td>
</tr>
<tr>
<td>Digit or Number</td>
<td>0</td>
</tr>
<tr>
<td>Superscript</td>
<td>S</td>
</tr>
<tr>
<td>Roman numeral</td>
<td>R</td>
</tr>
<tr>
<td>Small letter</td>
<td>Y</td>
</tr>
<tr>
<td>Bullet</td>
<td>B</td>
</tr>
<tr>
<td>Inverted question</td>
<td>I</td>
</tr>
<tr>
<td>Symbols</td>
<td>Z</td>
</tr>
<tr>
<td>Special space</td>
<td>*</td>
</tr>
<tr>
<td>No replacement</td>
<td>?</td>
</tr>
</tbody>
</table>

We applied the above procedure on the training data set described in the Methods section and came up with a predefined list of “special” capital word patterns (Table 2).

A first word in a sentence could be classified as a “special” capital word or not by first represent it as a symbol-based
word using the replacement table (Table 1) and then compare the symbol-based word with the predefined list of “special” capital word patterns (Table 2).

The algorithm to detect a “special” capital word for the first word in a sentence starts by truncating the given word at its first uppercase letter and then represents the truncated word as a list of Unicode names where each Unicode name is associated with each character of the truncated word. Next, it converts a list of Unicode names into a symbol-based word using the replacement table (Table 1). Finally, it checks the symbol-based word against the predefined list of “special” capital word patterns (Table 2) to determine whether or not the given word is a “special” capital word.

Table 2. Predefined list of “special” capital word patterns

<table>
<thead>
<tr>
<th>Sentence First Words</th>
<th>“Special” Capital Word Patterns</th>
</tr>
</thead>
<tbody>
<tr>
<td>Start with digits</td>
<td>&quot;0H&quot;, &quot;0&quot;, &quot;0P0H&quot;, &quot;0PH&quot;, &quot;0HP0H&quot;, &quot;0HP&quot;, &quot;0P0PH&quot;, &quot;0POPH&quot;, &quot;0XH&quot;, &quot;0POPOP0H&quot;, &quot;0HPOP0H&quot;, &quot;0HXH&quot;, &quot;0HP0HP0H&quot;, &quot;0P0P0POP0PH&quot;, &quot;0HP0&quot;</td>
</tr>
<tr>
<td>Start with punctuations</td>
<td>&quot;P&quot;, &quot;H&quot;, &quot;PHPH&quot;, &quot;PH&quot;, &quot;PHP&quot;</td>
</tr>
<tr>
<td>Start with lowercase letter</td>
<td>&quot;XH&quot;, &quot;X0H&quot;, &quot;XP0PH&quot;</td>
</tr>
<tr>
<td>Start with symbols or others</td>
<td>&quot;Z&quot;, &quot;B&quot;, &quot;S&quot;, &quot;*&quot;, &quot;T&quot;, &quot;ZH&quot;, &quot;RP&quot;</td>
</tr>
</tbody>
</table>

2. The abbreviations detection algorithm: Based on our observations of MEDLINE citations in our training data set, we have found that one of the ways sentences are incorrectly segmented is due to the failure of recognizing abbreviations between consecutive sentences. To minimize the impact of this problem, we build a predefined list of abbreviations (Figures 1 and 2) from last words of sentences and develop a simple abbreviations detection algorithm using it.

Abbreviations could connect to sentences that have the first character of their first word as either a lowercase or uppercase letter; however, we limit the algorithm’s attention to just lowercase letters and consider only abbreviations having 8 characters or less to reduce the number of collected samples.

To build a list of abbreviations, we collected all words that (a) have 8 characters or less, (b) are ended with a period, and (c) are followed with a word starting with a lowercase letter. We then tabulated the collection of the collected words into two groups: 1-period group (e.g., “approx.”, “Corp.”, etc.) and 2-period group (e.g., “i.e.”, “Ph.D.”, etc.) and filtered out words with a frequency of less than 100 to build the predefined list of abbreviations.

We applied the above procedure on the training data set and came up with the following list of abbreviations for the 1-period group (Figure 1) and the 2-period group (Figure 2).

Figure 1: The “general” 1-period abbreviation group

Figure 2: The “general” 2-period abbreviation group

The simple abbreviations detection algorithm could be implemented by checking the sentence last word. If the sentence last word without trailing whitespace characters is found in the above 1-period or 2-period abbreviation group, then it is an abbreviation.

**System Overview**

The algorithm starts with segmenting a document into sentences with each SBD engine, followed by a majority voting to build a list of most voted sentence “proposals” for the document. After this, the content-based features and the grammar-based features are computed using the five sentence-based POS tagging types, the abbreviation detection algorithm, the capital word detection algorithm, and computing general sentence statistics. These features are then used to determine whether to connect consecutive sentences. The algorithm ends with splitting sentences with specific patterns.

To decide on the connection among segmented sentences, the algorithmic approach is to look for complete and incomplete sentences. Incomplete sentences could be the results of SBD engines that failed to recognize sentence delimiters, identify capital words, detect abbreviations, etc. Further, incomplete sentences should connect to other sentences.
We used sentence attributes, the number of votes on each sentence, sentence structure, and POS to evaluate the completeness of each segmented sentence. There are two sets of features for each sentence: content-based and grammar-based. Content-based features are based on character statistics, abbreviations, majority voting, and five sentence-based POS tagging types. Grammar-based features are based on the POS tags for the first and last two terms in a sentence. These two sets of features complement each other to help in making connection decisions between consecutive segmented sentences. Figure 3 shows the workflow of the MEDLINE SBD system to process a MEDLINE citation and each step of this process is described in detail in the next section.

Figure 3: MEDLINE Sentence Boundary Detection Diagram

**MEDLINE Sentence Boundary Detection Process**

The MEDLINE SBD process consists of six steps: (1) segment a document into sentences by each SBD engine, (2) build voted sentences among SBD engines using a majority voting algorithm, (3) generate the content-based features for voted sentences, (4) generate the grammar-based features for voted sentences, (5) apply the rules-based SBD algorithm to decide on connection between consecutive sentences, and (6) split sentences with specific patterns.

1. **Segment a document into sentences:** Each SBD engine segments the document into sentences and each sentence is recorded with a starting index and a stopping index relative to the start of the document and its voting SBD engine.

2. **Build voted sentences using a majority voting algorithm:** The majority voting algorithm groups sentences from SBD engines together and assigns the number of votes for each sentence to build the voted sentences for the document. For the three SBD engines, the algorithm collects sentences into three groups: 3-vote group, 2-vote group, and 1-vote group. First, sentences with 3 votes are collected into the 3-vote group. Next, sentences with 2 votes that are not in the 3-vote group are collected into the 2-vote group. Finally, sentences with 1 vote that are not in the 3-vote and 2-vote groups are collected into the 1-vote group.

Note that only one SBD engine will be selected for the last group with one vote to avoid overlapping among selected
3. Generate the sentence content-based features: The content-based features consist of structure information and character statistics of a sentence, and they are used to justify if a sentence is complete or incomplete for the purpose of connecting consecutive sentences. There are six sentence content-based features: “GRAMMAR”, “STYLE”, “ABBREVIATION”, “LENGTH”, “LOCATION”, and “CATEGORY”, and they are defined and explained next.

A sentence is complete when it begins with a capital letter (a regular/special capital word), ends with a punctuation mark (period, question mark, or exclamation point), and has both a subject and a verb. A subject may be a noun (a person, place, or thing) or a pronoun. For example, a sentence that starts with a capital word and consists of one or several combinations of “NOUN-Phrase sentence” and “VERB-Phrase sentence” has a high chance to be a complete sentence. As a result, we used the five sentence-based POS tagging types and the “special” capital word detection algorithm defined in the Methods section (Basic definitions and Algorithms to Handle Sentence Boundaries Challenges in MEDLINE) to generate the “GRAMMAR” feature and the “STYLE” feature, respectively.

| **GRAMMAR** = “Grammatical Complete” | If the POS tagging type = NP+VP+NP or NP+VP or VP+NP |
| **GRAMMAR** = “Grammatical Incomplete” | If the POS tagging type = NP or VP or Unknown |

| **STYLE** = “Upper-case Sentence” | If a sentence has a “regular” or “special” capital word |
| **STYLE** = “Lower-case Sentence” | Otherwise |

Next, we apply the abbreviations detection algorithm defined in the Methods section (Algorithms to Handle Sentence Boundaries Challenges in MEDLINE) on the last word of the sentence to build the “ABBREVIATION” feature.

| **ABBREVIATION** = “Yes” | If an abbreviation is found at the end of the sentence |
| **ABBREVIATION** = “No” | Otherwise |

Based on our observations, a sentence with an average sentence length (AVGSL ≈ 110 characters) or higher is most likely to be a complete sentence. In addition, the higher number of votes assigned on a sentence from three SBD engines gives more confidence that it is a complete sentence. Therefore, the number of characters and the number of votes are used to define a sentence content-based feature called “LENGTH” as follows:

| **LENGTH** = “Medium-or-Long Sentence” | If (votes <= 2 and number of characters >= ¼ of AVGSL) or (votes > 2 and number of characters >= ½ of AVGSL) |
| **LENGTH** = “Short Sentence” | Otherwise |

The location of a sentence is another feature, and its values include below:

| **LOCATION** = “First” | If the sentence is the first sentence |
| **LOCATION** = “Last” | If the sentence is the last sentence |
| **LOCATION** = “Middle” | Otherwise |

In addition, the punctuations are used to define one more feature: “CATEGORY”.

| **CATEGORY** = “Punctuations-Only-Sentence” | If the sentence consists of punctuations only |
| **CATEGORY** = “Regular” | Otherwise |

4. Generate the sentence grammar-based features: POS categorizes words by type, such as nouns, verbs, adjectives, and each sentence is associated with a POS list that is used to build the grammar-based features. As shown in the Methods section (Basic definitions), some combinations between the ending POS terms of the current sentence and the beginning POS terms of the next sentence could help to connect them together. Therefore, we define two grammar-based features for each sentence: “BEGINNING POST TERMS (BPT)” and “ENDING POST TERMS (EPT)”. The first feature is based on the first one or two POS terms of the sentence while the second one is based on the last one or two POS terms of the sentence.

The four values of the BPT are “normal-POS”, “PUNCT+normal-POS”, “special-POS”, and “PUNCT+special-POS”. 

| **normal-POS** | The POS list starts with a normal-POS |
| **PUNCT+normal-POS** | The POS list starts with a PUNCT followed with a normal-POS |
| **special-POS** | The POS list starts with a special-POS (“PROPN” or “INTJ”) |
| **PUNCT+special-POS** | The POS list starts with a PUNCT followed with a special-POS |
The four values of the EPT are “normal-POS”, “normal-POS+PUNCT”, “special-POS”, and “special-POS+PUNCT”.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>normal-POS</td>
<td>The POS list ends with a normal-POS</td>
</tr>
<tr>
<td>normal-POS+PUNCT</td>
<td>The POS list ends with a normal-POS followed with a PUNCT</td>
</tr>
<tr>
<td>special-POS</td>
<td>The POS list ends with a special-POS</td>
</tr>
<tr>
<td>special-POS+PUNCT</td>
<td>The POS list ends with a special-POS followed with a PUNCT</td>
</tr>
</tbody>
</table>

For example, consider the following sentence with its POS list:

Sentence: “This method shows advantages in two aspects.”
POS list: ['DET', 'NOUN', 'VERB', 'NOUN', 'ADP', 'NUM', 'NOUN', 'PUNCT']

The BPT feature is “normal-POS” because its POS list starts with a normal-POS (DET) and the EPT feature is “normal-POS+PUNCT” because its POS list ends with a normal-POS and a PUNCT (NOUN, PUNCT).

5. Apply the rules-based SBD algorithm: We apply the rules-based SBD algorithm to connect or separate segmented sentences and the following statements are used in the algorithm.

- Set the sentence “START” or “STOP” means it is the start sentence or the end sentence, respectively.
- Set the sentence “RIGHT CONNECT” means it connects to the next sentence on its right.
- Set the sentence “LEFT CONNECT” means it connects to the previous sentence on its left.

There are heuristic rules and three parts of observation rules, and they are executed in this order. In addition, there are synchronization rules to synchronize connections.

5a) Heuristic Rules: There are eight straightforward heuristic rules, and they are created to handle common-sense cases. For example, sentences ended with abbreviations should be “RIGHT CONNECT” to the next sentence.

<table>
<thead>
<tr>
<th>Rule</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>If sentence CATEGORY = “Punctuations-Only-Sentence” then If sentence LOCATION = “First” then set the sentence to “RIGHT CONNECT” Else if sentence LOCATION = “Last” then set the sentence to “LEFT CONNECT” and “STOP”</td>
</tr>
<tr>
<td>2</td>
<td>If sentence LOCATION = “First” then set the sentence to “START”</td>
</tr>
<tr>
<td>3</td>
<td>If sentence LOCATION = “Last” then set the sentence to “STOP”</td>
</tr>
<tr>
<td>4</td>
<td>If sentence LENGTH = “Medium-or-Long Sentence” then set the sentence to “START”</td>
</tr>
<tr>
<td>5</td>
<td>If sentence ABBREVIATION = “Yes” If sentence LOCATION = “First” or “Middle” then set the sentence to “RIGHT CONNECT”</td>
</tr>
<tr>
<td>6</td>
<td>If the sentence starts with a case-insensitive “Copyright” or a symbol® then If the sentence contains a year between 1900 and 2200 then set the sentence to “START”</td>
</tr>
<tr>
<td>7</td>
<td>If sentence LENGTH = “Short Sentence” and GRAMMAR = “Grammatical Complete” then If sentence STYLE = “Uppercase Sentence” then set the sentence to “START”</td>
</tr>
<tr>
<td>8</td>
<td>If sentence LENGTH = “Short Sentence” and GRAMMAR = “Grammatical Incomplete” and If number of words &lt;= 1 or number of words without stop words &lt;= 1 or number of characters &lt;= 5 then If sentence LOCATION = “First” then set the sentence to “RIGHT CONNECT” Else if sentence LOCATION = “Last” then set the sentence to “LEFT CONNECT”</td>
</tr>
</tbody>
</table>

5b) Synchronization rules: These rules are straightforward rules, and they are required to synchronize connections among surrounding sentences when there is a new connection between sentences.

<table>
<thead>
<tr>
<th>Rule</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>If the current sentence is not the first sentence: If it is set to “LEFT CONNECT” then set the previous sentence to “RIGHT CONNECT” Else if it is set to “START” then set the previous sentence to “STOP”</td>
</tr>
<tr>
<td>2</td>
<td>If the current sentence is not the last sentence: If it is set to “RIGHT CONNECT” then set the next sentence to “LEFT CONNECT” Else if it is set to “STOP” then set the next sentence to “START”</td>
</tr>
</tbody>
</table>
5c) Observation rules - Part 1: When a sentence is a short sentence and does not start with a capital word, it should connect to its preceding sentence as described in the following two rules.

Rule 1: If the current sentence GRAMMAR = “Grammatical Incomplete” and STYLE = “Lowercase Sentence” and LENGTH = “Short Sentence” then
Set the current to “LEFT CONNECT” and the preceding to “RIGHT CONNECT”

Rule 2: If the current sentence GRAMMAR = “Grammatical Complete” and STYLE = “Lowercase Sentence” and LENGTH = “Short Sentence” then
If the preceding sentence GRAMMAR = “Grammatical Incomplete” and LENGTH = “Short Sentence” then
Set the current to “LEFT CONNECT” and the preceding to “RIGHT CONNECT”

5d) Observation rules - Part 2: Rules developed in this section are more complicated than the rest since both content-based features and grammar-based features of the current and next sentences are involved to make the connection decisions. There are many cases where sentences are broken into pieces due to errors in recognizing sentence delimiters, capital words, and abbreviations; most of these sentences are incomplete. There are twelve rules shown in Figure 4 that are created based on the combinations between sentence content-based features GRAMMAR, STYLE, LENGTH, and sentence grammar-based features EPT and BPT and they are applicable only to “Short” sentences.

<table>
<thead>
<tr>
<th>RULE</th>
<th>CURRENT SENTENCE</th>
<th>NEXT SENTENCE</th>
<th>ACTIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Grammatical Incomplete</td>
<td>Uppercase</td>
<td>normal-POS or normal-POS+PUNCT or special-POS or special-POS+PUNCT</td>
</tr>
<tr>
<td>2</td>
<td>Lowercase</td>
<td>Short</td>
<td>normal-POS or normal-POS+PUNCT or special-POS or special-POS+PUNCT</td>
</tr>
<tr>
<td>3</td>
<td>Grammatical Incomplete</td>
<td>Uppercase</td>
<td>Short</td>
</tr>
<tr>
<td>4</td>
<td>Grammatical Incomplete</td>
<td>Uppercase</td>
<td>Short</td>
</tr>
<tr>
<td>5</td>
<td>Grammatical Incomplete</td>
<td>Uppercase</td>
<td>Short</td>
</tr>
<tr>
<td>6</td>
<td>Grammatical Incomplete</td>
<td>Lowercase</td>
<td>Short</td>
</tr>
<tr>
<td>7</td>
<td>Grammatical Incomplete</td>
<td>Lowercase</td>
<td>Short</td>
</tr>
<tr>
<td>8</td>
<td>Grammatical Incomplete</td>
<td>Lowercase</td>
<td>Short</td>
</tr>
<tr>
<td>9</td>
<td>Grammatical Incomplete</td>
<td>Lowercase</td>
<td>Short</td>
</tr>
<tr>
<td>10</td>
<td>Grammatical Incomplete</td>
<td>Lowercase</td>
<td>Short</td>
</tr>
<tr>
<td>11</td>
<td>Grammatical Incomplete</td>
<td>Lowercase</td>
<td>Short</td>
</tr>
<tr>
<td>12</td>
<td>Grammatical Incomplete</td>
<td>Lowercase</td>
<td>Short</td>
</tr>
</tbody>
</table>

Figure 4: Observation rules - Part 2

5e) Observation rules - Part 3: The following rule is applicable to short and incomplete sentences and the connection decision is based solely on the sentence grammar-based feature BPT of the next sentence.

Rule 1: If the current sentence GRAMMAR = “Grammatical Incomplete” and STYLE = “Lowercase Sentence” and LENGTH = “Short Sentence” then
If the next sentence GRAMMAR = “Grammatical Incomplete” and LENGTH = “Short Sentence” then
If the next sentence BPT = “normal-POS, PUNCT+normal-POS, special-POS, or PUNCT+special-POS” then
Set the current to “RIGHT CONNECT” and the next to “LEFT CONNECT”

684
6. Split sentences with specific patterns: Through our observations, we identified some specific patterns that SBD engines failed to recognize, and they are as follows:

<table>
<thead>
<tr>
<th>Pattern #1</th>
<th>Pattern #2</th>
</tr>
</thead>
<tbody>
<tr>
<td>It starts with a space, an uppercase letter, a period, a space, and a capital word such as “B. The” or “A. In”</td>
<td>It starts with a word, a period, and a capital word such as “test. The” or “station. That”</td>
</tr>
</tbody>
</table>

For both patterns, we observed that when sentences in the left and right of the period are grammatically complete (GRAMMAR = “Grammatical Complete”) and the POS term of the first word followed the period is ‘DET’, ‘ADV’, ‘ADJ’, ‘VERB’, ‘PRON’ or ‘NUM’, sentences could be split at the period.

Therefore, the method to split sentences with specific patterns could start with the detection of the pattern #1 or #2 in a sentence and follow by splitting the sentence into the left sentence and the right sentence at the period. When both sentences are grammatically complete and the POS term of the first word of the right sentence belongs to the above-mentioned POS list, the sentence splitting is acceptable.

Discussion

We implemented the SBD algorithm and conducted experiments on MEDLINE citations randomly selected from several different biomedical journals. The algorithm is a combination between a majority voting among three SBD engines and rule-based post-processing algorithms. Some rules are heuristic, while others are derived from observations and from processing the collected training data set. The SBD algorithm is trained from the 2020 MEDLINE training data set and evaluated from two ground-truth data sets and these data sets are described in detail in the Methods section.

For the evaluation, we measure the SBD performance in terms of accuracy using two criteria: sentence-based and citation-based. For the sentence-based, the accuracy is based on the number of sentences detected by an SBD engine that are found in the ground-truth data set (Tables 3A and 3B). For the citation-based, the accuracy is relied on the number of citations having all their sentences found in the ground-truth data set (Tables 4A and 4B).

<table>
<thead>
<tr>
<th>Table 3A: The GENIA corpus sentence-based evaluation results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total GENIA corpus sentences: 18,541</td>
</tr>
<tr>
<td>Python NLTK</td>
</tr>
<tr>
<td>Number of sentences found in the GENIA data set:</td>
</tr>
<tr>
<td>18,036</td>
</tr>
<tr>
<td>Accuracy:</td>
</tr>
<tr>
<td>97.28</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 3B: The GENIA corpus citation-based evaluation results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total GENIA corpus citations: 1,999</td>
</tr>
<tr>
<td>Python NLTK</td>
</tr>
<tr>
<td>Citations with all sentences found in the GENIA data set:</td>
</tr>
<tr>
<td>1,761</td>
</tr>
<tr>
<td>Accuracy:</td>
</tr>
<tr>
<td>88.09</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 4A: The 2021 SBD sentence-based evaluation results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total 2021 SBD sentences: 11,204</td>
</tr>
<tr>
<td>Python NLTK</td>
</tr>
<tr>
<td>Number of sentences found in the 2021 SBD data set:</td>
</tr>
<tr>
<td>10,899</td>
</tr>
<tr>
<td>Accuracy:</td>
</tr>
<tr>
<td>97.28</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 4B: The 2021 SBD citation-based evaluation results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total 2021 SBD citations: 1,020</td>
</tr>
<tr>
<td>Python NLTK</td>
</tr>
<tr>
<td>Citations with all sentences found in the 2021 data set:</td>
</tr>
<tr>
<td>885</td>
</tr>
<tr>
<td>Accuracy:</td>
</tr>
<tr>
<td>86.76</td>
</tr>
</tbody>
</table>

The average time in milliseconds to process a citation on a “DELL Intel(R) Core(TM)i7-8700 CPU @ 3.20GHz with 64.0 GB running 64-bit OS Windows 10 Enterprise” are: 1.21 ms, 2.22 ms, 1.36 ms, and 6.27 ms for Python NLTK, pySBD, Syntok, and MEDLINE SBD, respectively.

The results show that the combined MEDLINE SBD engine offers the best sentence boundary detection accuracy on the ground-truth data sets compared to the other three single SBD engines.
Conclusion

We developed the MEDLINE SBD algorithm using majority voting among multiple SBD engines and post-processing rules derived from content-based and grammar-based sentence features. Evaluation results show that our proposed SBD algorithm yielded good performance in terms of accuracy when compared to each single SBD. Therefore, we conclude that the MEDLINE SBD algorithm provides an improved approach to sentence segmentation for MEDLINE citations. However, we identified two shortcomings that should be resolved to further improve the detection results. First, the algorithm starts with a voting from sentences segmented by each SBD engine and follows with rules to decide whether to connect consecutive sentences. Even though there were some brute-force parsing operations in the last step of the MEDLINE SBD process to split segmented sentences with certain specific patterns, these operations actually just help to minimize the impact of the problem by handling obvious cases, but the problem still remained to be solved. Additionally, when all SBD engines do not agree on detected sentence boundaries, a predefined single engine is selected for the task and the algorithm cannot select other engines that might have better sentence boundaries. As a result, we plan to expand the current approach by including sophisticated methods to split segmented sentences as well as techniques to dynamically select an appropriate single engine when none of the engines agree on detected sentence boundaries.

Acknowledgments

This work was supported by the Intramural Research Program of the U.S. National Institutes of Health (NIH), National Library of Medicine (NLM), and Lister Hill National Center for Biomedical Communications (LHNCBC).

References

9. Gillick D. Sentence boundary detection and the problem with the u.s. NAACL Short ’09. 2009; 241–244.
17. Sentence segmentation and word tokenization (syntok) [internet]. Available from https://pypi.org/project/syntok/
Strategies for Disease Containment: A Biological-Behavioral-Intervention Computational Informatics Framework

Eva K Lee, PhD1,*, Yifan Liu, PhD 1, Fan Yuan, PhD 1, Ferdinand H Pietz, MPH2

1NSF-Whitaker Center for Operations Research in Medicine and Healthcare, Georgia Institute of Technology, Atlanta, GA; 2Strategic National Stockpile, U.S. Department of Health and Human Services, Washington DC.

Abstract

In this study, we describe the development and use of a biological-behavior-intervention computational informatics framework that combines disease modelling for infectious virus with stratifications for social behavior and employment, and resource logistics. The framework incorporates heterogeneous group behavior and interaction dynamics, and optimizes intervention and resources for effective containment. We demonstrate its usage by analyzing and optimizing containment strategies for the 2014-2016 West Africa Ebola outbreak, and its implementation for responses to the 2020 COVID-19 pandemic in the United States. Our analysis shows that timely action within 1.5 months from the onset of confirmed cases can cut down 90% of overall infections and bring rapid containment within 6-8 months. The additional medical resources required are minor and would ensure proper treatment and quarantine of patients while reducing the risk of infections among healthcare workers. The benefits of infection control (e.g., infection control / death control) would be reduced by 10 to over 100 fold and time to containment would increase by 2-4 fold when intervention and medical resources are injected within 5 months. In contrast, the additional resources needed to bring down the overall infection in a delayed intervention are significant, with inferior results. The disease module can be tailored for different pathogens. It expands the well-used SEIR model to include social and intervention activities, asymptomatic and post-recovery transmission, hospitalization, outcome of recovery, and funeral events. The model also examines the transmission rate of healthcare workers and allows for heterogenous infection factors among different groups. It also captures time-variant human behavior during the horizon of an outbreak. The framework optimizes the intervention timeline and resource allocation during an infectious disease outbreak and offers insights into resource availability in time and quantity can affect the disease trends and containment significantly. This can inform policy, disease management and resource allocation. While focusing on bed availability for quarantine and treatment appears to be simplistic, their necessity for Ebola responses cannot be overemphasized. We link these insights to a web-based tool to provide quick and intuitive observations for decision making and investigation of the disease outbreak situation. Subsequent use of the system to determine the optimal timing and effectiveness and tradeoffs analysis of various non-pharmaceutical intervention strategies for COVID-19 provide a foundation for policy makers to execute the first-step response. These results have been implemented on the ground since March 2020. The web-based tool pinpoints accurately the import of disease from global travels and associated disease spread and health burdens. This prospectively affirms the importance of such a real-time computational system, and its availability before onset of a pandemic.

*Corresponding author: evalee-gatech@pm.me

This work is partially supported by a grant from the Centers for Disease Control and Prevention and the National Science Foundation (1P1-P1516074). The findings and conclusions in this paper are those of the authors and do not necessarily represent the official position of the funding agencies.

1. Introduction

SARs, bird flu, H1N1, Ebola, Zika and the current SARS-CoV-2 outbreaks underscore the critical importance of emergency response and medical preparedness. Such needs are widespread as globalization and air transportation facilitate rapid disease spread across the world. For the past two decades, as the public health community worked to strengthen national systems to avoid international spread of disease, governing bodies increasingly recognized that biological threats not only have global health impacts but also wide-ranging socioeconomic disruptions.

Mathematical models of dynamic systems and their interactions are powerful tools for analyzing disease outbreaks and supporting real-time decision-making. Computational tools of infectious disease outbreaks and epidemics offer insights in propagation patterns and help policy makers to synthesize potential interventions. This facilitates strategic planning and establishment of guidelines to prevent further disease spread. These models focus on describing the mechanism of disease propagation and model it as processes that result in the transition of individuals of the population between different disease stages or compartments. While many existing models incorporate contact tracing to predict spread pattern, key elements on optimal usage of scarce resources and effective and efficient process performance (e.g., diagnostics and screening, non-pharmaceutical interventions (NPIs), trained personnel/robots and medical resources for treatment, decontamination, vaccination) have not been integrated.

Since Kermack and McKendrick (1) established the theory of the SIR models and other corresponding compartmental models in the 1920s and 1930s, ordinary-differential-equation based compartmental models are among the most popular methods. Researchers continue to expand and employ them in modeling numerous infectious diseases.
that vary according to the specific pathogen. It
disease progression and transmission (light blue)
example, for Ebola, safe funerals offer a venue to
pharmaceutical, in combination or alone, play
Interventions, pharmaceutical or non-
dynamics on how disease spread takes hold.
beige) to model heterogeneous transmi ssion
environmental influences, and risk factors (light
logistics (Figure 1). The framework accounts for
social behavior and employment,  and resource
modelling for infectious virus with stratifications
Our informatics framework combines disease
2 Methods and Design
Over the last two decades, our team has worked with thousands of public health leaders and frontline emergency responders, providing them with decision support tools and analytic-informatics expertise as they respond to numerous domestic and global public health crises, and establish state and federal guidelines. Specifically, we have developed
specialized mathematical theory and computational systems for numerous infectious diseases including smallpox, hepatitis A, H1N1, Ebola, Zika, MERS, avian flu, and COVID-19 (17,19,20,21,24-28) and have the honor to assist on-the-ground operations with great success (16,18). Our experiences and lessons learnt reveal the challenge and urgency to derive a unified mathematical theory that is applicable to any type of infectious disease (29) and develop a holistic biological-behavior-intervention computational framework that allows for in-depth validation and adoption.

In this study, we design a biological-behavior-intervention system computational decision modeling framework that simultaneously (i) captures disease spread characteristics, (ii) models social, human behavior and dynamic environment, (iii) incorporates resource usage and intervention logistics, and (iv) allows for system optimization to minimize infection and mortality under time and resource constraints. We first test our model using the 2014-2016 real data from Guinea, Sierra Leone, and Liberia and compare the optimal resource and predicted trend of the outbreak against the actual reported cases. Our result suggests that timely and optimal allocation of limited medical resources during pandemics is crucial to minimize the cumulative infections and achieve early containment. Using these historical data, the system predicts with promising results the recent 2018-2020 Ebola outbreak in DRC. We link these insights and design a web-based tool for public health and policy makers to investigate the disease transmission outbreak and intervention impact. The web-application is universal and practical and can be generalized to other infectious diseases.

Since December 2019, the system has been used for COVID-19 response. Using early results from Germany and China, we established human-to-human transmission and asymptomatic transmission in January 2020. In February 2020, we employed the system to optimize the timing and impact of non-pharmaceutical intervention (NPI) of school closure, business telework, facemask and social distancing. Our findings echo the Ebola results, and underscore the importance of early NPI implementation that can significantly reduce deaths and overall infections, and facilitate early containment. Our recommendations to local and state officials triggered the initial school closures. It also facilitated adoption of telework by a major private organization. Furthermore, the prospective prediction of confirmed cases and deaths for a collection of cities and counties prove to be very accurate. Currently, we continue to support safe re-opening, mass vaccination and evidence-based and personalized treatment.

2 Methods and Design
Our informatics framework combines disease modelling for infectious virus with stratifications for social behavior and employment, and resource logistics (Figure 1). The framework accounts for disease progression and transmission (light blue) that vary according to the specific pathogen. It incorporates human behavior, social and environmental influences, and risk factors (light beige) to model heterogeneous transmission dynamics on how disease spread takes hold. Interventions, pharmaceutical or non-pharmaceutical, in combination or alone, play crucial roles in containing the disease. For example, for Ebola, safe funerals offer a venue to minimize disease spread that would otherwise
occur during traditional funeral ceremonies. This is both human/social behavior change as well as a means of non-pharmaceutical intervention. Other behavioral changes such as handwashing and use of face masks can have significant positive impact in combating contagious infectious disease, as evidenced by the current COVID-19 pandemic (26-28).

2.1 Compartmental Model for Disease Propagation and Dynamic Heterogeneous Interaction

Our disease module employs an ordinary differential equation-based disease propagation model and generalizes the SEIR model to the specific pathogens. For Ebola, due to the high possibility of transmission during traditional funerals without protection, deceased patients who do not receive cremation will stay infectious. A new stage \( F \) for the funeral is added to model the transmission happens during funerals. If a deceased patient is buried instead of cremated, it is not removed from the system immediately; instead, it stays in stage \( F \) for a period of time before being removed. Another stage introduced to the model is \( H \) for hospitalization. Depending on available medical resource, infectious individuals can be hospitalized, and are quarantined from outside civilians. Therefore, patients in stage \( H \) are not infectious to civilians but are infectious to healthcare workers who have direct contact with them. Similar to stage \( I \), a patient in stage \( H \) can either recover (\( R \)) or die, and the deceased hospitalized patients can either be buried or cremated. Thus, the disease transmission system for civilians contains seven stages: \( S \) (susceptible), \( E \) (exposed), \( I \) (infectious), \( R \) (recovered), \( H \) (hospitalized), \( F \) (funeral), and \( D \) (deceased; removed from the system), as depicted in Figure 2a.

Healthcare workers are also in danger of getting infected and they face different situations than local civilians. Although they have proper protection within the hospital, there is still a possibility of getting infected. There are resources such that healthcare workers are dealt with caution to minimize cross infection to others, i.e., there is no funeral stage (of potential infection) for healthcare workers. Let \( HS, HE, HI, HH, HR, \) and \( HD \) denote the 6 stages for healthcare workers (Figure 2b).

These two groups are not isolated: Healthcare workers get exposed (\( HE \)) because of their close physical contact with patients who are hospitalized (\( H \)) as well as with infectious healthcare workers (\( HI \)). For brevity, in the results, healthcare workers are assumed to be better informed and protected than outside civilians, and hence do not get infected by contacting infectious civilians (\( I \)). Furthermore, if healthcare workers are infected (\( HI \)) they do not infect healthy civilians (\( S \)) as they take extra precautions. In section Discussion, we describe time-variant parameters that are introduced to handle varying transmission factors among heterogeneous groups.

The final component is the integration of the two compartmental models above with interrelation. We define the disease stage space \( \Phi = (S, E, I, R, H, F, D, HS, HE, HI, HH, HR, HD) \). Let the civilian population in the system who are not hospitalized be \( N = S + E + I + R + F \), and let the total number of healthcare workers be \( HN = HS + HE + HI + HR + HD \). Let \( \mu_s \) denote the average number of infectious contacts each individual makes in a unit time, i.e., the baseline contact rate; \( \mu_F \) is the transition rate from exposed stage to the infectious stage and it is characterized by the incubation period of the virus; \( \mu_I \) is the baseline transition rate from infectious stage to stage \( R, F, \) or \( D \) and it is determined by the duration of the infection; \( \mu_H \) is the transition rate from infectious stage to hospitalized or the inverse of the time between infection to admission to healthcare facility; \( \mu_H \) is the transition rate from stage \( H \) to stage \( R, F, \) or \( D \), or the inverse of the duration of hospitalization; \( \mu_F \) is the transition rate from stage \( F \) to stage \( D \) and it is characterized by the duration of the funeral.

Other parameters in the model include \( H_m \), the total number of beds available. Although Ebola is a severe, often fatal disease, getting medical care early can make a significant difference. This reflects the availability of healthcare resource in the West African countries during the outbreak: \( p_{HD} \) and \( p_D \) are the death rate due to infections with and without hospitalization, respectively. \( p_{HF} \) and \( p_F \) are the percentages of deceased patients who will go through traditional burial (non-cremation) with and without hospitalization, respectively. \( p_{SF} \) is the adjusted factor of baseline contact rate for funerals/unsafe burials, and \( f \) is the adjusted factor of healthcare worker transmission rate, reflecting their proper protection.

The following system represents the dynamics of their heterogeneous interaction:

\[
\begin{align*}
\frac{d}{dt} S &= -\mu_s S \left( 1 + p_{SF} F \right) \\
\frac{d}{dt} E &= \mu_s S \left( 1 + p_{SF} F \right) - \mu_E E \\
\frac{d}{dt} I &= \mu_E E - \mu_I I - \mu_I I \\
\frac{d}{dt} H &= \mu_I I - \mu_H H \\
\frac{d}{dt} HS &= -\mu_{HF} \frac{HS}{HN + HH} \left( H + HI + HH \right) \\
\frac{d}{dt} HE &= \mu_{HF} \frac{HS}{HN + HH} \left( H + HI + HH \right) - \mu_E HE \\
\frac{d}{dt} HI &= \mu_E HE - \mu_H HI \\
\frac{d}{dt} HH &= \mu_H HI - \mu_H HH
\end{align*}
\]
The model divides the disease spread into two categories: (i) disease propagation outside healthcare facilities that infects civilians, and (ii) disease propagation within healthcare facilities that infects healthcare workers. In addition, the heterogeneous infectivity among civilians and healthcare workers is characterized by factor \( f \) to capture the proper knowledge and protections of healthcare workers. Numerous risk groups can be expanded based on different disease vulnerabilities and transmission vectors / factors (19,40). The modeling framework is not limited to Ebola outbreaks or hospital beds; by setting proper parameters and types of intervention, it can be used to estimate and predict the potential disease trend of outbreaks of other infectious diseases and strategize the best intervention options. The Results section includes tailored model for SARS-CoV-2 analyses.

2.2 Optimization Framework for Intervention, Resource Allocation, and Disease Containment

One major challenge faced by resource-constrained countries is the lack of healthcare facilities and infrastructures. The current COVID-19 pandemic clearly underscores the devastation and the intense healthcare and resource burden. This is sorely felt in developing countries. During the 2014 Ebola outbreak, Guinea did not have enough hospital beds or healthcare facilities, thus not all patients could be properly quarantined or receive medical care. Without available beds, patients could only be treated at home, which brought risks to their relatives and neighbors. The parameter \( p_H \) in the model can be used to capture the available beds (yes, or no): if there are beds available at hospitals or healthcare facilities, then \( p_H = 1 \), i.e., infectious patients will be admitted to quarantine; otherwise \( p_H = 0 \). Essentially, \( p_H \) can be seen as an indicator of whether there are vacancies for quarantine and treatment. Therefore, \( p_H = \{ H < H_m \} \), where \( I \) is the indicator function. Thus \( p_H \) denotes if an infected patient receives the needed medical resources. We use the number of beds available in each affected location as the main factor in controlling the spread of the virus and investigate the impact of availability and delayed supply of beds; and determine the number of beds required to contain the outbreak. We estimate total infections and the timeline of the outbreak, including those for healthcare workers. These additional beds and the associated resources and healthcare workers are from international assistance. However, international assistance is limited, and devoting too many resources to an infected area does not necessarily accelerate the containment, rather it may increase risk of infections among healthcare workers. Therefore, we design an optimization framework to minimize the number of beds required and the associated medical resources subject to the constraints on the total infections among civilians and healthcare workers at containment and the time needed to achieve containment. In particular, the following constrained optimization problem will be solved numerically:

\[
\min_{H_m} \quad I \leq I_0, H \leq H_0, T \leq T_0
\]

such that:

\[
\frac{d}{dt} R = (1 - p_D) \mu I + (1 - p_H) \mu_H
\]

\[
\frac{d}{dt} F = \mu_H \mu I + \mu_F F
\]

\[
\frac{d}{dt} D = (1 - p_F) \mu I + (1 - p_H) \mu_H F + \mu_F F
\]

\[
\frac{d}{dt} HR = (1 - p_H) \mu_H HH
\]

\[
\frac{d}{dt} HD = p_H \mu_H HH
\]

\[
\frac{d}{dt} H = p_H \mu_H HH - \mu_F F
\]

\[
\frac{d}{dt} I = \mu_H F - \mu I - (1 - p_F) \mu I - (1 - p_H) \mu_H I
\]

\[
\frac{d}{dt} F = \mu_F F - \mu_F F
\]

\[
\frac{d}{dt} D = (1 - p_F) \mu I + (1 - p_H) \mu_H D + \mu_F F
\]

\[
\frac{d}{dt} H = p_H \mu_H HH - \mu_F F
\]

where \( I \) and \( H \) are the numbers of infections among civilians and healthcare workers by the end of outbreak, respectively, while \( I_0 \) and \( H_0 \) being the target number of infections among civilians and healthcare workers at the end of the outbreak. Here, \( T_0 \) is the target time of containment, and \( T \) is the actual time of containment, and defined as \( T = \min \{ t : \frac{1}{I_0} t e^{-\ln 2} \leq \delta \} \), i.e., the earliest time such that the change rate of infections is controlled by a certain level \( \delta \). Other constraints can also be introduced to obtain the optimal allocation of healthcare resources under different scenarios (e.g., labor, protective gears, ventilators etc). The variables in this optimization framework are the output of the system of ordinary differential equations of the disease module, thus searching for optimal solutions is computationally expensive. We apply a line search algorithm for a given structure of parameters and search for the optimal setting iteratively on the surface generated by the solution to the differential equations. To avoid local minima, the line search is conducted on multiple starting points. Each iteration requires about 10 CPU minutes to solve.

3. Results

3.1 2014-2016 Ebola Numerical experiments are first performed for the 2014-2016 outbreak in Guinea, Sierra Leone, and Liberia to validate the model and predict the trend of disease spread. Table 1 summarizes the common parameters used by the experiments. Other parameters are country-specific, in particular the basic reproduction number is estimated using the CDC historical statistics (30). Table 2 summarizes the individualized parameters. For each country, we use the first five months for training and parameter estimates, and then we predict the disease trends. We optimize the timing and resource needs to achieve disease containment within 6 months and contrast the outcome against several scenarios.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value (reference)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Incubation period (1/(\mu_I))</td>
<td>Uniform distribution between 8 and 12 days (31,32)</td>
</tr>
<tr>
<td>Infection duration (1/(\mu_I))</td>
<td>Uniform distribution between 5 and 9 days (32,33)</td>
</tr>
<tr>
<td>Hospitalization duration (1/(\mu_H))</td>
<td>Uniform distribution between 5 and 15 days (32)</td>
</tr>
<tr>
<td>Funeral duration (1/(\mu_F))</td>
<td>1 day (28)</td>
</tr>
</tbody>
</table>
WHO report, Guinea had 160 beds by October 250 total infections (blue curve). According to the 200 additional beds were made available by disease outbreak. The optimal (minimum) resource requirements for containment within 6 months occurs when 50 beds observed starting from April 15, 2014. The basic reproduction number by mid 2015 with approximately 3,800 infections 2014. With 160 beds, the epidemic was contained by the end of 2015 (red solid curve). This reflects closely the real situation (orange dotted curve). If the intervention reflect ing the change of human behavior and the effort of interventions with public awareness of the disease outbreak. The optimal (minimum) resource requirements for containment within 6 months occurs when 50 beds are available within a month. The epidemic could have been rapidly contained with no more than 250 total infections (blue curve). According to the WHO report, Guinea had 160 beds by October 2014. With 160 beds, the epidemic was contained by mid 2015 with approximately 3,800 infections by the end of 2015 (red solid curve). This reflects closely the real situation (orange dotted curve). If 200 additional beds were made available by November 2014, the outbreak would have been contained by the end of 2015 with roughly 3,500 cumulative infections (yellow curve). We observe that timing is critical. Rapid containment can be achieved with minimal extra resources if they become available early on. Late availability of beds only marginally reduces the overall infections.

The analyses for Sierra Leone and Liberia bear similar patterns. Figure 4 shows the performance of the 3 countries under different resources and response timelines evaluated under six metrics: time to containment, resource needs, total infections and deaths, healthcare workers (HCW) infections and deaths. We normalize the results with respect to the real data. Hence, a value 1 means it performs the same as the real situation, < 1 is an improvement, and > 1 is a degradation. The brown curve uses the resources reported by WHO. Our model results in metrics all within the interval (0.95, 1.06), indicating that the model predicts well. Uniformly, the results show that timely action within 1.5 months (optimal: blue curves) from the onset of confirmed cases can cut down 90% of overall infections and bring rapid containment within 6-8 months. The optimal medical resources (beds) required are minor (16.6% to 31.3% with respect to the real-world situation) and would ensure proper treatment and quarantine of patients while reducing the risk of infections among healthcare workers. If intervention of medical resources are injected within 5 months (non-optimal), infections will be reduced by over 50% with containment achieved within 12 months. However, when compared to the optimal solution, its benefit (in infection/death control) would be reduced by 10 to over 100 fold and time to containment would increase by 2-4 fold when intervention and medical resources are delayed. We illustrate one such non-optimal solution (Liberia, deep blue curves) that uses 10 times the bed resources, and results in 24-fold

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Guinea</th>
<th>Sierra Leone</th>
<th>Liberia</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic reproduction number ($R_0$)</td>
<td>1.45</td>
<td>1.4</td>
<td>1.45</td>
</tr>
<tr>
<td>From infection to admission ($1/\mu_H$)</td>
<td>3-4 days (32)</td>
<td>4 days (3 days after Oct 1, 2014) (32)</td>
<td>4 days (3 days after Oct 1, 2014) (32)</td>
</tr>
<tr>
<td>Number of Healthcare Workers as of Oct 1, 2014</td>
<td>300 (36)</td>
<td>1000 (36)</td>
<td>500 (36)</td>
</tr>
</tbody>
</table>

Table 2. Parameters Specific for Each Country

Figure 3 shows the number of predicted infections in Guinea under different scenarios. The intervention effect was observed starting from April 15, 2014. The basic reproduction number $R_0$ changes twice (1.45 to 0.9, then to 1.3) during the intervention reflecting the change of human behavior and the effort of interventions with public awareness of the disease outbreak. The optimal (minimum) resource requirements for containment within 6 months occurs when 50 beds are available within a month. The epidemic could have been rapidly contained with no more than 250 total infections (blue curve). According to the WHO report, Guinea had 160 beds by October 2014. With 160 beds, the epidemic was contained by mid 2015 with approximately 3,800 infections by the end of 2015 (red solid curve). This reflects closely the real situation (orange dotted curve). If 200 additional beds were made available by November 2014, the outbreak would have been contained by the end of 2015 with roughly 3,500 cumulative infections (yellow curve). We observe that timing is critical. Rapid containment can be achieved with minimal extra resources if they become available early on. Late availability of beds only marginally reduces the overall infections.

The analyses for Sierra Leone and Liberia bear similar patterns. Figure 4 shows the performance of the 3 countries under different resources and response timelines evaluated under six metrics: time to containment, resource needs, total infections and deaths, healthcare workers (HCW) infections and deaths. We normalize the results with respect to the real data. Hence, a value 1 means it performs the same as the real situation, < 1 is an improvement, and > 1 is a degradation. The brown curve uses the resources reported by WHO. Our model results in metrics all within the interval (0.95, 1.06), indicating that the model predicts well. Uniformly, the results show that timely action within 1.5 months (optimal: blue curves) from the onset of confirmed cases can cut down 90% of overall infections and bring rapid containment within 6-8 months. The optimal medical resources (beds) required are minor (16.6% to 31.3% with respect to the real-world situation) and would ensure proper treatment and quarantine of patients while reducing the risk of infections among healthcare workers. If intervention of medical resources are injected within 5 months (non-optimal), infections will be reduced by over 50% with containment achieved within 12 months. However, when compared to the optimal solution, its benefit (in infection/death control) would be reduced by 10 to over 100 fold and time to containment would increase by 2-4 fold when intervention and medical resources are delayed. We illustrate one such non-optimal solution (Liberia, deep blue curves) that uses 10 times the bed resources, and results in 24-fold
total infection and 2.2-fold in time to containment with respect to the optimal solution. However, it still outperforms the real situation: using 80% of the resources, reducing 25% of the time to containment and 41%-55% infection and deaths. The orange curves correspond to delayed interventions that use excess amounts of medical resources, yet the results are inferior to the optimal solution, and a marginal improvement over the real-world situation. These findings reaffirm that with delayed intervention, the extra resources needed to bring down the overall infections are significant.

At the onset of disease, the spread appears to be extremely sensitive to timing of available beds. The red curve on Liberia shows that even with the same resources (as the brown curve), a mere delay of 2 days in receiving half of the resources would result in 23% increase in infections and deaths Bed resources available at the fast-spreading phase of the outbreak has a huge impact on its propagation. Policy makers must act decisively, and the impact can be very significant.

Our analyses also reveal that early intervention has critical impact on healthcare workers. Although protected by protective gear against the virus, healthcare workers are still at high risk of infection due to their close contact with infected patients. At the initial phase of the pandemic, i.e., the second half of 2014, the number of infections among healthcare workers increased drastically. For example, according to our model, under the real-world bed availability, the infections among healthcare workers in Liberia reached 250 after 6 months of the initial outbreak and 368 at disease containment. There are several reasons to explain this phenomenon: insufficient personal protective gear before international intervention, improper use of protective gear, and severe shortage of medical staff among these resource-constrained countries to combat a relatively large outbreak. We note that the loss of medical staff including experienced doctors and nurses made it increasingly difficult to control the outbreak. This triggered the necessary assistance by foreign countries, including sending in healthcare workers who could provide intensive patient care.

3.2 COVID-19 In January 2020, we expanded our 6-stage SEPAIR model (17) that incorporates pre-symptomatic infectious (P) and asymptomatic infectious (A) to include post-recovery infectious (Q), recovered-immune (R), and recovered-not-immune (T) to model the SARS-CoV-2 virus (Figure 5). Below, we briefly present our findings regarding timing and effectiveness of NPI during the early stage of the pandemic.

Based on public data from China and Germany, we established the first analytic results of SARS-CoV-2 virus and the resulting 2019 coronavirus disease. Specifically, our analyses reveal the index case in China must have begun no later than November 15, 2019; and that asymptomatic spread contributes substantially (at least 1/3 to ½) to the disease propagation. These critical insights set a tone of urgency as we discussed our findings to state and local public health emergency leaders. Subsequent analysis in February 2020 confirmed the effectiveness of NPI in Singapore and Hong Kong. This along with public data from Diamond Princess, allowed us to model NPI interventions in dozens of cities and counties in the United States in February 2020, including New York City, King County Seattle, Santa Clara and San Diego California, Washington DC, Atlanta Georgia, and Montgomery, Baltimore counties in Maryland (26-28).

Table 3 summarizes the parameters used in the SARS-CoV-2 analyses. Other parameters are region-specific, including population, total number of students, workforce / labor statistics, and travel patterns. Figure 6 shows that uniformly across the U.S., timely school closure (upon confirmed onset of infected cases) can potentially reduce 32%-77% of the total infection within 4 weeks, and 98% within 12 weeks. A delay of two weeks would render only 1-27% reduction in the initial month, and 94% by 14 weeks. Adding 50% of tele-work workforce can reduce infection over 77% within 4 weeks and 97% within 8 weeks. Without NPI, and with the current bed availability, about 48% - 85% of population will be infected before herd immunity can be achieved. It also shows that with a compliance of 90%, facemasks can reduce 35%-57% of infection. It illustrates timing of implementation is of paramount importance, and that without these interventions, our healthcare system will be overwhelmed.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Assumption</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean exposed duration</td>
<td>2 days (25%), 5.2 days (55%), and 12.5 days (20%)</td>
<td>(37)</td>
</tr>
<tr>
<td>Mean pre-symptomatic infectious duration</td>
<td>1 day</td>
<td>Estimate (38)</td>
</tr>
<tr>
<td>Mean symptomatic, asymptomatic durations</td>
<td>3 days</td>
<td>Estimate (38)</td>
</tr>
<tr>
<td>Mean post-recovery infectious probability and duration</td>
<td>3 days</td>
<td>Estimate (38)</td>
</tr>
<tr>
<td>Symptomatic probability</td>
<td>2/3</td>
<td>Assumption (17)</td>
</tr>
<tr>
<td>Mean time between showing symptoms to quarantine/hospitalization</td>
<td>4.5–9 days</td>
<td>(37)</td>
</tr>
<tr>
<td>Mean hospitalization time</td>
<td>9.2 days</td>
<td>(37)</td>
</tr>
</tbody>
</table>
closed schools and implemented various NPI including size of group gatherings and face mask usage. The results also triggered Microsoft to cancel 3 leadership conferences (involving thousands of business leaders from around the world) and promptly initiated telework for their workforce \((39)\).

### Table 4. Prospective infection and mortality prediction through July 31 2020 obtained from our model in February 2020. NPI interventions include school closure and 50% business telework, with remaining business practicing social distancing, self-protection, staggering work shifts and protection of all essential workers. The range represents various compliance rates (70%-90%) of facemask and social distancing. Bold are the reported cases and deaths.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Infections</td>
<td>Deaths</td>
<td>Infections</td>
</tr>
<tr>
<td><strong>New York City</strong></td>
<td>1,673 – 3,021</td>
<td>38 – 155</td>
<td>62,750 – 113,197</td>
</tr>
<tr>
<td><strong>Washington DC</strong></td>
<td>136 – 281</td>
<td>4 – 8</td>
<td>1,200 – 2,485</td>
</tr>
</tbody>
</table>

3.3 A web-based tool for investigating disease transmission, intervention strategies, and aiding decision-making We design a web-based tool, RealOpt-Ebola©, on top of our existing CDC emergency response decision support system RealOpt-Regional© \((14-16,18,20)\) (Figure 7). RealOpt-Ebola links to our computational framework and is tailored to help public health and policy makers investigate the transmission of a disease outbreak (both regionally and internationally). The system translates user input into backend mathematical models. Users can analyze scenarios and strategize interventions to control the outbreak. Since solving the ODE disease module is computationally expensive on the client side, we design a modified fast algorithm of the inclined decay with an exponential adjustment model \((5)\). The fast heuristics returns solution rapidly with minimal difference in solution quality. Different scenarios analysis can be performed by specifying different model parameters (e.g., the average number of contacts per infected individual each day, risk factor, and the distribution of incubation period, etc.). Users can also introduce disease control by adjusting the control effectiveness. The system then simulates the interventions being implemented and returns the containment curves showing how the disease spread would have changed as a result.

RealOpt-Ebola© includes all recorded Ebola cases occurring since 2014. Users can select any month and view the number and geographic distribution of the cases, and perform disease spread and predictive analysis. Besides modeling and visualizing the Ebola outbreak (e.g., in West Africa, DRC), the interactive map tool in RealOpt-Ebola can model the influence of the outbreak to other countries (e.g., infected patients arrived via international flights). Specifically, when performing the predicted infection trend in a country, e.g., Africa, some infected individuals may travel to other countries outside by air. In this case, users can initialize a new scenario analysis in the destination of that flight and evaluate the impact of the traveling patients. A new disease model will be initialized which simulates the disease spread scenario at the destination. If there are no travel restrictions, infectious travelers can lead to disastrous consequences. RealOpt-Ebola’s surveillance and simulation modules enable emergency planners to analyze the future trend of the outbreak with an up-to-date information feed and evaluate the strategies for containment and the severity of disease spread due to the infectious travelers. Supported by proper model parameters (from historical data or estimates), this tool can also be used to monitor other regional disease outbreaks.
Since 2016, we have disseminated RealOpt-Ebola on the ground in Africa, assisting in biosurveillance, monitoring, and tracking of risk and potential spread, and rapid response actions. Policy makers use the predicted outcome to produce actionable policies regarding preparedness timeline and expected risks. The system is also adapted to RealOpt-COVID to help predict and evaluate impacts of global travel restrictions in the early stage of the COVID-19 outbreak.

4. Discussion
This work focuses on designing a holistic biological-behavior-intervention system computational decision modeling framework that simultaneously (i) captures disease spread characteristics, (ii) models social, human behavior and dynamic environment, (iii) incorporates resource usage and intervention logistics, and (iv) allows for system optimization to minimize infection and mortality under time and resource constraints. To the best of our knowledge, the modeling framework is the first that incorporates biology of the virus, human and social behavior and available resources, and intervention and operations logistics into a single computational framework. The system allows population-based analysis as well as county-level local studies.

Analysis and modeling provide important tools to understanding the usefulness of timely interventions and risks associated with delays. The outcomes presented herein offer an argument for having capacity for timely interventions.

The disease module adapts to different pathogens and disease pathways. It allows heterogeneous compartmental modules for different (risk) groups and with dynamic contact environments. Human behaviors that can be modelled include those that reduce disease spread such as self-protection, and those that fuel spread such as crowded dwellings and work environments, and certain social events or religious / cultural beliefs. All these play key roles in the disease dynamic. Resource needs arise naturally -- beds, providers, personal protective equipment, drugs, medical and life-support devices, vaccines, etc. These interlay with operations processes and supply-and-demand logistics. Smart and rapid resource allocation must be carried out under limited time and resources when minutes mean life-and-death.

The informatics framework allows one to focus on the problem at hand and analyze optimal strategies, whether it is to optimize timing and bed allocations for the treatment of patients, or optimal timing in rolling out combinations of NPIs. All these decisions require actions by policy makers. The informatics framework empowers policy makers to take decisive actions with informed risks and tradeoffs and understanding of the look-ahead situations.

During a pandemic or any critical mission, time, labor, and resources are essential yet are limited. Policy makers face tough decisions on actions, when and how to carry out, and allocate optimally what types of resources, how many and to whom, with the goals to protect the population and save as many lives as possible. The optimization module can accommodate various objectives. In this paper, we demonstrate its usage in optimizing the timing for intervention and (minimizing) the resources needed to contain the Ebola crisis; and the timing of combination strategies of non-pharmaceutical interventions that minimizes total infections and deaths for COVID-19. The system considers individual behavior, personal actions, and social influence. It can also be used, among a wealth of applications, to optimize resource allocation in vaccine prioritization, mass diagnostic testing, and vaccination events (14,15,16,18,19,40). The interlace of optimization within simulation and ODE systems is novel and useful yet computationally challenging.

For the 2014-2016 Ebola outbreak in Guinea, Sierra Leone, and Liberia, using the initial infection on the ground and the initial 5 months as training, our model predicts the subsequent disease trends given the available medical resources. The optimal timing and resource needs to ensure rapid disease containment within 6 – 8 months can be achieved through early intervention (within 1.5 months of onset of confirmed cases). In this case, a relatively small resource investment yields over 90% reduction in overall infections and deaths. Our Ebola predictions, reported first in November 2014 (24) and January 2015 (25) to the State Department and the United Nations, was promising and rather accurate in reflecting the on-the-ground situations. In 2016, the team was in W. Africa training local healthcare and emergency responders in using the early version of the RealOpt-Ebola webtool for monitoring and optimizing their resources. Predictions for the 2018-2020 DRC Ebola outbreak showcase that historical data can be highly informative in predicting future trends, empowering policy makers to take timely risk-informed actions.

Since December 2019, our system has been used for modeling and evaluating effectiveness and timing of intervention strategies for COVID-19, with some results implemented on the ground in the United States since March 2020 (26-28,40). Using data from other countries, the predictions for the US offer insightful risks and tradeoffs. The NPI intervention and case/death prediction was accurate and was adopted by policymakers during the initial stage of the
pandemic. Our analysis shows that early NPI intervention can save lives (98%) and achieve disease containment within 6-9 months. Delaying actions add significant disease and resource burden while producing inferior overall outcome. The web-based tool pinpoints accurately the import of disease from global travels and associated disease spread and health burdens. This prospectively affirms the importance of such a real-time computational system, and its availability before onset of a pandemic.

COVID-19 has revealed major challenges in science-based recommendations. The rippling and interdepending downstream impacts of NPI create social, economic, and political implications that prove to be extremely controversial for timely decisions. Nonetheless, analytic results offer actionable steps and opportunities for valuable discussion by policy makers. Our theoretical and computational framework continues to facilitate states and local jurisdictions, including determining optimal diagnostic and pooling strategies, vaccination prioritization, mass vaccination, variant effects on herd immunity, and the after-COVID next generation health systems delivery.

The model can handle heterogeneous transmission and risk factors. If the transition rate from stage $I$ to stage $H$ is faster, which requires rapid diagnosis of infectious patients and immediate transition to hospitalization, the outbreak would also be contained much faster. The infections among healthcare workers in affected populations are also estimated. Healthcare workers remain at high risk of getting infected due to lack of training, insufficient protective equipment and working in highly contagious and stressed conditions.

The informatics framework is general-purpose and allows analyses of complicated scenarios, such as multiple groups with different parameters. In this paper, the baseline contact rate between civilians is treated as a linear relationship. Since virus can spread via different means and contacts, the possibility of infection rate of individuals in stage $S$ may be nonlinear depending on how they interact with the infected ones, or whether they make direct contact, and under what environment. Figure 8 shows types of contacts of different risk groups and environments. Under this setup, the civilians, healthcare workers, event handlers, etc. can be further divided into groups with different parameters associated with different contacts. Similarly, as more organizations get involved in the outbreak, other groups such as volunteers and military forces can be added to the model with their own compartments and model parameters. This enables us to predict the number of infections among different groups and evaluates their risk to mount respective mitigation strategies, respectively. In our vaccine prioritization and mass vaccination work, populations are divided into different risk-groups according to age, health conditions and employment nature, and vaccine allocation and timing are optimized. The model simultaneously tracks infections/deaths and optimizes overall disease spread for both intra-facility and various indoor activities, and outside in the public, where dynamic transmission patterns are incorporated (19,40).

The equation captures the behavior of healthcare workers during different times of the day. Other equations in the system parameter a constant value, we set the parameters as functions of time. During different times of a day (different social events, or different phases of the pandemic), people interact differently, and this affects the spread of the virus. For example, healthcare workers have different infection rates depending on if they are inside their work facilities or at home or in social settings. If a healthcare worker is infected but not hospitalized yet, civilians may get infected if their social contact with infected civilians, or attend an unsafe funeral, or social events. Then

\[
\frac{d}{dt} S = -\mu_S \left( I + p_F S + \eta(t) p_H I \right)
\]

where $\eta(t)$ is a time-variant parameter indicating if the healthcare workers are outside the healthcare facilities at time $t$, and $p_D$ is the proportion of these off-duty healthcare workers. Similarly, off-duty healthcare workers may come into contact with infected civilians, or attend an unsafe funeral, or social events. Then

\[
\frac{d}{dt} HS = -\left( 1 - \eta(t) \right) \mu_F \frac{HS}{HN + HH} (H + H + HH) - \eta(t) \mu_F \frac{HS}{(1 - p_D)HN + HH} (H + (1 - p_D)H + HH) + \frac{p_D HS}{p_D HN + N} (I + p_F S + p_H I)
\]

This equation captures the behavior of healthcare workers during different times of the day. Other equations in the system can be updated accordingly. Parameters such as the basic reproduction rate can also be defined as functions of time to
accommodate the changes in policy or environment during the outbreak. This enables the modeling and analysis of complicated situations.

In our recent work, we apply these heterogeneous risk factors and dynamic transmissions to optimize vaccine prioritization and mass vaccination strategies. Our findings show that it is of paramount importance to split available vaccines optimally among the high-risk groups and the general public (40). The reduction in total infections and deaths, as shown in Figure 9, is significant.

The biological-behavior intervention modeling and optimization framework developed herein can be adapted to other kinds of epidemics. With proper set up of the compartmental model and parameters, as well as the definition of interventions, and medical / resources and the role they play in mitigating the spread of the disease, one can utilize the system and analysis procedure to predict the future development of an outbreak and optimize the timing, and types (and combinations) of interventions and optimal allocation of resources to contain it. This will be of great significance in the current and future global pandemic preparedness.

References

4. Lee EK, Liu Y, Pietz FH. A compartmental model for Zika virus with dynamic human and vector populations
Incidence and Impact of Missing Functional Elements on Information Comprehension using Audio and Text

Gondy Leroy, PhD\textsuperscript{1}, David Kauchak, PhD\textsuperscript{2}, Nicholas Kloehn, PhD\textsuperscript{1}

\textsuperscript{1}University of Arizona, Tucson, AZ, USA; \textsuperscript{2}Pomona College, Claremont, CA, USA

Abstract

Audio is increasingly used to communicate health information. Initial evaluations have shown it to be an effective means with many features that can be optimized. This study focuses on missing functional elements: words that relate concepts in a sentence but are often excluded for brevity. They are not easily recognizable without linguistics expertise but can be detected algorithmically. Two studies showed that they are common and affect comprehension. A corpus statistics study with medical (Cochrane sentences, N=44,488) and general text (English and Simple English Wikipedia sentences, N=318,056 each) showed that functional elements were missing in 20-30\% of sentences. A user study with Cochrane (N=50) and Wikipedia (N=50) paragraphs in text and audio format showed that more missing functional elements increased perceived difficulty of reading text, with the effect less pronounced with audio, and increased actual difficulty of both written and audio information with less information recalled with more missing elements.

Introduction

Health literacy is vital for achieving and maintaining good health. In the previous decade in the US, several national programs have emphasized this goal and its importance. For example, the Affordable Care Act\textsuperscript{(1)} emphasized patient-centeredness, the National Action Plan to Improve Health Literacy\textsuperscript{(2)} specified national goals, and the Plain Writing Act\textsuperscript{(3)} demanded clarity in government communications\textsuperscript{(4)}. However, these existing guidelines do not focus and have not been validated for new information distribution methods. They focus on text and visual presentation, but to our knowledge, they have not been updated for audio.

As technology evolves, new modes for communication are being created to provide access to health information. Audio is becoming increasingly popular with mobile devices using virtual assistants and smart speakers. Smart speakers are now ubiquitous in US households and are increasingly used for health-related application. By 2020 there were about 87.7M smart speakers used in the US\textsuperscript{(5, 6)} with annual growth of 30-40\%. Hospitals are testing the use of smart speakers\textsuperscript{(7, 8)} and in 2019, about 16\% of all questions asked of a smart speaker were health and wellness questions\textsuperscript{(6)}. There are few studies and little existing expertise on how to create audio or to combine audio with text. For example, the Plain Language Summit in 2019\textsuperscript{(9)}, described plain language as “language that people can understand the first time they read or hear it,” but audio was not covered during the summit.

With the increasing popularity of audio as an information source, research is needed to design the best strategies for using this resource as a medium for disseminating health information. In our early work, we focused on discovering lexical, semantic, syntactic, and composition features in text with demonstrated impact on comprehension. We are expanding this work to audio to discover how best to create content for audio consumption. We have already established that health information consumers can learn from audio information as well as from text information\textsuperscript{(10)}. Critically, unlike previous work that only focused on text, we now examine both text and audio to see how text features affect comprehension in the different presentation formats.

Sentences in English rely upon syntactic structure to convey the meaning of a sentence. The structure of a sentence dictates the order of words in a sentence, but also relies on function words that help join different phrases and clauses in a sentence. These function words tell us about the relationships between content words, but often carry little content themselves. For example, the word ‘that’ in a phrase such as ‘parasites that impact’, relates the two content words. Often, these function words are left out in spoken language or in complex written language for various reasons. When these words are left out, i.e., missing functional elements, the person reading the sentence has less information available to decode the meaning of the sentences than if all of these words were included.

We hypothesize that when there are missing functional elements in a sentence, the sentence will be more difficult to readers than if these functional elements were included because explicit connections between content-bearing terms are omitted. We developed an algorithm to detect missing functional elements in text. Using this algorithm, we examine the frequency of missing function elements in two corpora (medical and general) and evaluated the effect of missing elements on comprehension with both text and audio. Missing elements are a very frequent phenomenon,
occurring in almost a third of the sentences in our corpora. We found that there is a subtle effect of missing functional elements on comprehension and we found differences between the text and audio presentation. Even though the effects are subtle, because missing elements occur so frequently, the effect is important.

The current work brings two contributions. The first is a demonstration that some text features, not easily recognize by non-experts, are frequently found in text, hence the usefulness of algorithmically identifying them. The second contribution is in showing how such features affect comprehension and how this differs depending on the mode of interaction with the information (text vs audio). For brevity, we will use the phrase ‘missing elements’.

**Background**

**Content Simplification.** Most literacy research has focused on text as the mode of distribution and examined how to make that text as accessible as possible. Given some existing text, the goal of text simplification is to transform that text into a variant that is more understandable. In the healthcare community, the majority of work has used readability formulas to aid content creators (e.g., medical writers) in simplifying text. A variety of different metrics have been proposed, e.g., Flesh-Kincaid and Reading Ease, and Simple Measure of Gobbledygook (SMOG), and some more comprehensive tools, e.g., Coh-Metrix, have been created. The scores assigned by these metrics and tools indicate the difficulty level of a text and there are agreed upon limits for text intended for use by non-experts. For example, the advice is to write at 6th to 8th grade level. This approach suffers from three main problems. First, there is a single score assigned to an entire text; unless a writer is very familiar with the inner workings of the formula, adjusting the text to improve the score is difficult. Second, these formulas rely on simple text statistics that tend to correlate with text readability, e.g., word length, but often don’t accurately measure, on a per instance case, the actual text difficulty. Well written, understandable text tends to have better readability score, however, simply manipulating the components of the readability metrics does not create text that is more readable or easier to understand. The following examples were evaluated using the built-in Flesh-Kincaid formulas in Microsoft Word:

- Eating fruit is a healthy habit and enjoyable too. (7.5 Grade Level)
- Eating fruit is. A healthy habit and enjoyable too. (5.8 Grade Level)
- Eating fruit is. A healthy habit. And enjoyable too. (5.2 Grade Level)

Our own work and that of others have shown that there is no relationship between these scores and reader comprehension. Third, as we consider new modes of communication, i.e., audio, these formulas have not been validated and it is not clear how the formulas will translate to this new medium.

On the other end of the spectrum, there has been a lot of work recently that focuses on fully automated approaches to text simplification that do not require any human editing. These approaches usually rely on a sentence-aligned training corpus of difficult sentences aligned to corresponding simple sentences. Most of these approaches have leveraged models that have been successful for machine translation, including phrasal models and syntactic models and, most recently, sequence-to-sequence neural networks. There are two key challenges with applying these models in the health and medical domain. First, the performance of these models, including our own, is still not nearly good enough for practical application, particularly in a domain like health where correctness is critical. Second, currently, there are no good large sentence-aligned corpora of medical text available. While general domain text will provide some guidance, it is important to have domain text to capture both the lexical and syntactic nuances.

In this work, we take a compromise between these two extremes and view the text simplification processes as a semi-automated approach where a human in the loop utilizes a tool to help guide the writing process. Unlike readability formulas that only provide high-level information, the tool provides concrete guidance on which portions of the text are problematic and, critically, also provides concrete suggestions for how these sections can be improved. We have successfully used a range of data sources, corpus statistics, and machine learning to discover many features and develop individual algorithms evaluated through user studies. The algorithms that are shown to impact readability and understandability are integrated in an online text editor. The current version of the tool contains a range of tool components at the word level, sentence level, and corpus level. In this work, we are expanding to focus on features that are useful for audio content.

---

1 http://simple.cs.pomona.edu:3000/
Importance of Audio Information. While text is still a common mode for providing information in medicine, a new approach is emerging fast: audio information presentation. An important contributing factor is the rapidly rising use of virtual assistants and smart speakers. In 2018, 30% of Americans used voice to find and purchase products. Smart speakers have become an increasingly common household item and offer a range of activities with an increasing number of interactions that focus on healthcare. For example, in 2021 there were more than 2,000 skills in the Alexa’s Health and Fitness section, several focus on tracking (medication, menstruation, fertility, calories), finding providers and scheduling, but a large portion is devoted to providing information (e.g., drug information, general health advice, and information on specific conditions and treatments). A 2019 survey found that 52% of those surveyed would welcome using a voice assistant for healthcare and many are already asking information about symptoms (73%), medication (46%), hospitals (38%), research treatment (38%), nutrition (39%) and more. The COVID pandemic has sped up several initiatives using text or audio. For example, Microsoft and the Centers for Disease Control and Prevention (CDC) collaborated on a COVID-19 chatbot and the World Health Organization (WHO) released a chatbot using WhatsApp to provide relevant information.

Creating audio information from existing text is straightforward through the use of existing APIs, e.g., the Bing Speech Text-to-Speech API or Google’s Cloud Text-to-Speech API. Using audio itself is not new and has been used with visually impaired audiences. Example research has focused on e.g., effectiveness of audio instructions or dental programs for visually impaired children. Newer studies focusing on virtual assistants are often limited to evaluating the usefulness and correctness of answers in response to queries, e.g., a comparison of virtual assistant and their acceptance at home and usefulness in answering queries such as about gynecologic oncology.

However, there are increasingly more options for more sophisticated information exchanges, for example, Amazon Alexa competitions focusing on Type 2 diabetes support, hospitals providing Alexa skills or adding smart speakers in patient rooms (e.g., Cedars’ Sinai pilot study), and the increasing use of the Internet of Medical Things with voice-activated devices.

Our work differs from others in that we evaluate how to improve content for general, automated audio generation, a feature provided by virtual assistants although currently using existing, and not optimized, content. Since we focus on medical and healthcare content, the standards for accuracy are high and limited information (preferably none) can be omitted.

Methods

Identifying Missing Elements. We algorithmically identify missing elements in text based on the syntactic structure of the sentences. The algorithm does not require any human intervention and can be applied efficiently to text in a wide range of domains. Using this algorithm, we analyzed the frequency of missing elements in two different corpora. Then, we performed a user study to see the impact of missing elements on comprehension in both text and audio presentation.

Missing elements can be introduced into sentences in English in a variety of ways. In this work, we focus on two different ways that can occur in noun phrases. The first is missing elements in complex noun phrases. More specifically, we focus on noun phrases that contain a relative clause or verbal element in a nominal without an overt subordinate clause. For example, in the phrase

“The patients included are tested for ....”

there is syntactic information missing. This is often done in an attempt to be succinct, though it does leave some information implied rather than explicitly stated. A version of the sentence that includes this information could be written as:

“The patients who were included are tested for ...”

These missing elements can be detected using the grammatical structure of the sentence. A missing element occurs if there exists a noun phrase that contains a verb but does not contain a preposition or subordinating conjunction (i.e., a word with part-of-speech ‘IN’). We denote this Rule 1.

Another type of missing element is found in noun phrases that do not contain nouns, more specifically, having a nominal that lacks a determiner. In English, these include deictic determiners: function words that point or refer back to items that are not explicitly stated in the sentence. For example, in the phrase

3 https://cloud.google.com/text-to-speech/
“The patients included are tested for the disease, this indicates that... “

the term ‘this’ refers to a noun that is implied by the content but not stated. It can be made explicit, for example:

“The patients included are tested for the disease, this fact indicates that... “

A missing element of this type occurs if there exists a noun phrase that does not contain a noun or other words that act like nouns, i.e. a preposition, an existential ‘there’, a number, a gerund verb, or a comparative or superlative adjective. We denote this Rule 2.

For both rules, we can determine if a missing element occurs in a sentence, given the syntactic parse of the sentence. To apply the rules, we first preprocess the text with the Stanford CoreNLP toolkit to split the text into sentences and automatically parse the sentences. The rules conditions above are checked based on these automatically generated parse trees. In our case, we implemented the rules in Java. Rule 1 identifies as containing missing element all noun phrases that contain a verb (identified by one of the five verb part of speech tags), but that do not contain a preposition or subordinating conjunction (i.e., does not contain a word with the ‘IN’ part of speech tag). Rule 2 identifies as containing a missing element all noun phrases that do not contain a noun or a noun-like word, specifically, one of ten parts of speech (four noun tags, two adjective tags, a preposition, a number, existential there, or a gerund verb).

Evaluating Occurrence: Corpus Statistics Study. To get an understanding of the impact of these two types of missing elements we completed corpus analyses to evaluate whether these missing elements appear regularly in text and whether they are especially indicative of difficult text. We used two different corpora to measure the impact of missing elements. The first corpus is a medical corpus with abstracts collected from Cochrane articles. Cochrane provides a collection of a wide range of medical articles that are meant to “inform healthcare decision-making”. They represent digests of current research and represent a source where patients can obtain current information on diseases, treatments, and other medically-related information. We queried the database for 15 different illnesses that were the leading causes of death according to the CDC (e.g., heart disease) along with 4 common conditions (e.g., obesity) and collected the abstracts for all of the articles returned. The total dataset contains 44,488 sentences.

For the second corpus, we examined a corpus that contains 318,056 aligned sentence pairs from English Wikipedia and Simple English Wikipedia articles. Wikipedia articles are a common source for people searching for a wide range of topics and are written to be generally accessible. Simple English Wikipedia articles contain similar topics to the normal English Wikipedia articles, but written to be more accessible and broadly digestible. For each English Wikipedia sentence the corpus contains a corresponding sentence in Simple English Wikipedia with roughly equivalent information, though expressed more simply. This parallel corpus allows for a direct comparison of text with varying difficulty levels that is agnostic of topic. The Wikipedia dataset contains a broad range of topics and is an extremely commonly used resource. As noted above, currently no corpus exists of such sentences in the medical domain.

Evaluating Effects on Listeners: User Study. To evaluate the impact on reading and listening, we combined the data from two user studies where comprehension was tested with online consumers using a corpus of text snippets. The difference between the two studies was in how comprehension of the content was measured: with either multiple-choice questions or with free recall of information. These initial studies showed that consumer can learn new information from audio as well as from text which is a requirement before commencing research on optimizing content. The study was reviewed and approved by the Institutional Review Board of the University of Arizona.

The results for the current study are based on an analysis of our consumer comprehension data using the new missing element algorithms. The algorithms to detect missing elements (Rules 1 and 2) as well as the analysis of the relationship between missing elements and comprehension on the text snippets are new for this study and not reported elsewhere.

The goal was to compare user comprehension after being shown medical information two times. This simulates the situation where a patient is presented information after a medical consultation (e.g., orally in the doctor’s office and then again at home with written information). The studies were done in two phases designed to isolate the effect on comprehension of listening to information versus reading the same information. In the text-text variant, the information was presented twice to the participants as text. In the audio-text variant, the information was presented

4 https://www.cochranelibrary.com/
first as audio and then again as text. We used a total of 100 text snippets randomly selected from two medical sources and chosen to be of approximately equal length: Cochrane (N=50) and English Wikipedia snippets for health-related pages (N=50). For this paper, we calculated the number of missing elements based on Rules 1 and 2 in the text snippets to evaluate how the number of missing elements relates to comprehension.

In the first study, we used multiple-choice questions to measure comprehension. The participants were presented with the snippet (in either text form or audio form, depending on the condition) and then answered a multiple-choice question about the content (Multiple-Choice 1). They also scored the perceived difficulty of the snippet using a 5-point Likert scale with a score of 1 indicating ‘Very Easy’ text and 5 indicating ‘Very Difficult’ text. The participants were then shown the information again as text and they were given an opportunity to correct their answer to the same multiple-choice question (Multiple-Choice 2). We included the same question to see how much an individual improved with the second presentation of the information. For each multiple-choice question, there were three answer choices, one of which was correct. We measured comprehension with the percentage correct.

In the second study, we used free recall to measure comprehension; the study was identical to the first except that the first multiple-choice question was replaced with a free recall question where participants were asked to write, in their own words, as many pieces of information that they could remember about the information presented. To evaluate the quality of the free recall answers, we quantified how much of the information they recalled was in the original text that they were presented. To calculate this we first applied an automated spelling checker (to avoid differences due to participants who used a spelling checker and those who did not). Then, we measured the number of words that overlapped between the original text and the information recalled by the study participants.

To better understand the relationship between the information recalled and the original text, we used two measures of recall: exact and semantic. Exact recall was calculated by counting the number of content bearing words (i.e., nouns, verbs, adjectives, and adverbs) in the information recalled that occurred in the original text. This type of exact match can be too strict and does not allow for synonyms or other paraphrasing that participants might do. To capture a broader sense of overlap, we also used semantic recall, which is the number of content bearing words that were either an exact match of or that had a similar meaning to a word in the original text. To calculate if two words were similar, words were first represented by their word embedding. We then used the cosine similarity between word embeddings to identify word with similar semantic meaning. We used Google’s pre-trained 300-dimension word embeddings(37) and a cosine similarity threshold of 0.45 to tag words as semantically similar (this threshold was empirically determined).

Both studies were conducted using Amazon Mechanical Work (AMT) with three workers participating for each text snippet for each condition. Comprehension scores were averaged across the three workers and across the text snippets. We also report here on the time spent on the task, which is provided by AMT. Additional details of the AMT study have been reported in the study which focused on the impact of text-text and audio-text ordering on comprehension (38).

**Results**

**Corpus Statistics Study Results.** Table 1 shows the results of the corpus analysis for the two corpora. For each resource, we counted each occurrence of a missing functional element as well as the number of sentences containing one or more missing element. Since there can be multiple occurrence of missing elements in a single sentence (e.g., in different noun phrases), the number of occurrences can be higher than the number of sentences.

Missing elements occur frequently in medical text. The Cochrane dataset contained 15,734 instances of Rule 1 elements and almost a quarter of all sentences (23.5%) had at least one occurrence. Rule 2 was less frequent with 4,784 instances overall and almost a tenth of the sentences (9.7%) had at least one occurrence. Combining the rules, there was at least one missing element in 29.5% of the sentences.

Missing elements also tended to occur more frequently in difficult text, particularly Rule 1 elements. In the Wikipedia corpus 20.0% (63,737 sentences) of the English Wikipedia sentences contained a missing element versus 18.9% (60,067 sentences) for Simple English Wikipedia. This difference was almost entirely due to Rule 1 occurrences (41,545 sentences vs. 37,427 sentences, English vs. Simple English).
Table 1. The number and proportion of missing elements in the medical corpus and the English-Simple English sentence-aligned corpus.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Rule 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>occurrences</td>
<td>15,734</td>
<td>52,486</td>
<td>45,379</td>
</tr>
<tr>
<td>sentences</td>
<td>10,449 (23.5%)</td>
<td>41,545 (13.1%)</td>
<td>37,427 (11.7%)</td>
</tr>
<tr>
<td>Rule 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>occurrences</td>
<td>4,784</td>
<td>31,258</td>
<td>30,806</td>
</tr>
<tr>
<td>sentences</td>
<td>4,333 (9.7%)</td>
<td>27,479 (8.6%)</td>
<td>27,446 (8.6%)</td>
</tr>
<tr>
<td>Rule 1 or 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sentences</td>
<td>13,130 (29.5%)</td>
<td>63,737 (20.0%)</td>
<td>60,067 (18.9%)</td>
</tr>
</tbody>
</table>

User Study Results. Table 2 provides an overview of the study corpus used to measure user comprehension. Overall, missing elements are common in both types of text. The snippets contained on average 4.5 sentences and most contained one or more missing elements. There were more missing elements identified by Rule 1 (0.82 and 1.16 per snippet for Wikipedia and Cochrane snippets, respectively) than for Rule 2 (0.18 and 0.20 for Wikipedia and Cochrane snippets, respectively). The differences were not statistically significant.

Table 2. Corpus statistics for the text snippets used in the user comprehension study.

<table>
<thead>
<tr>
<th>Text Origin</th>
<th>Wikipedia</th>
<th>Cochrane</th>
</tr>
</thead>
<tbody>
<tr>
<td>Snippets (N)</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>Average Sentences / Snippet</td>
<td>4.50</td>
<td>4.52</td>
</tr>
<tr>
<td>Average Words / Snippet</td>
<td>94.36</td>
<td>96.12</td>
</tr>
<tr>
<td>Rule 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Minimum</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Maximum</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Average</td>
<td>0.82</td>
<td>1.16</td>
</tr>
<tr>
<td>Rule 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Minimum</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Maximum</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Average</td>
<td>0.18</td>
<td>0.20</td>
</tr>
</tbody>
</table>

In Table 3, we show the relationship with reading the text and the number of missing elements (text-text condition). We present reading first as it serves as a baseline: reading information is a familiar activity. We conducted
correlation analyses to estimate the impact of missing elements on the difficulty (actual or perceived) of text. We calculated one-tailed Pearson Correlation Coefficients using the average scores of the three AMT participants for each of the snippets (N=100). We chose one-tailed because we hypothesize that an increasing number of missing elements makes the text more difficult. We found that more time is spent on the tasks with more missing elements. The text is also perceived as more difficult when there are more missing elements. The effects are stronger for Rule 1 (which occurs more frequently in the text). There was no significant correlation with actual difficulty as measured here with multiple-choice questions.

Table 3. Results for the multiple-choice questions for the reading (text-text) condition. One-tailed Pearson correlation significant at 0.05 level* or at 0.01 level**.

<table>
<thead>
<tr>
<th>Missing Elements According to Rule Type</th>
<th>Text Difficulty Measure (N=100)</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Time Spent</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rule 1</td>
<td>.491**</td>
<td>.062</td>
<td>.068</td>
<td>.187*</td>
<td></td>
</tr>
<tr>
<td>Rule 2</td>
<td>.190*</td>
<td>.030</td>
<td>.346</td>
<td>.112</td>
<td></td>
</tr>
</tbody>
</table>

Table 4 shows the results for listening to the content first and reading afterwards (audio-text condition). When participants have little control over the time spent (they could not pause the recording) there is no relationship with time spent. Interestingly, we found a negative correlation with the perceived difficulty of the audio and the number of missing elements: when listening to the information, more missing elements make the information sound easier. While including functional elements may help make the connections more explicit in the sentences, leaving them out can sometimes help with the fluency of the sentences. Similar to reading text (above), there is no correlation between the actual difficulty, measured with the multiple-choice questions and the number of missing elements.

Table 4. Results for the multiple-choice for the listening (audio-text) condition. One-tailed Pearson correlation significant at 0.05 level* or at 0.01 level**.

<table>
<thead>
<tr>
<th>Missing Elements According to Rule Type</th>
<th>Text Difficulty Measure (N=100)</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Time Spent</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rule 1</td>
<td>.335</td>
<td>.065</td>
<td>.044</td>
<td>-.209*</td>
<td></td>
</tr>
<tr>
<td>Rule 2</td>
<td>.209</td>
<td>-.033</td>
<td>-.142</td>
<td>-.113</td>
<td></td>
</tr>
</tbody>
</table>

Our second study repeated this approach but replaced the multiple-choice questions with a free recall measure. This was done to provide a more sensitive measure of information comprehension and retention. Free recall was calculated as the overlap of words between the given content and the participants’ recall. Similar to the analysis above, we calculated the correlations between missing elements and the retention of information.

Table 5 shows the results with the free recall of information and reading text (text-text condition). There is no correlation between time spent and missing elements. While the number of exact or semantically matching words is not correlated to the missing elements (not surprisingly since many AMT workers prefer to work fast and spent a limited amount of time on a task), the overall similarity of the answer to the original information is strongly correlated with more missing elements resulting in a lower overall answer similarity, i.e., a worse answer by the participants. The information is perceived as more difficult when there are more missing elements.

Table 5. Results for the free recall measure for the reading (text-text) condition. One-tailed Pearson correlation significant at 0.05 level* or at 0.01 level**.

<table>
<thead>
<tr>
<th>Missing Elements According to Rule Type</th>
<th>Text Difficulty Measure (N=100)</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Time Spent</td>
<td>Exact Free Recall Count</td>
<td>Semantic Free Recall Count</td>
<td>Overall Recall Similarity</td>
<td>Perceived Difficulty</td>
</tr>
<tr>
<td>Rule 1</td>
<td>.044</td>
<td>-.077</td>
<td>-.060</td>
<td>-.268**</td>
<td>.180*</td>
</tr>
<tr>
<td>Rule 2</td>
<td>.127</td>
<td>.083</td>
<td>.086</td>
<td>-.123</td>
<td>.125</td>
</tr>
</tbody>
</table>
Table 6 shows the results for free recall of information after listening to the information first (audio-text condition). As expected, since participants have no control over audio play, there was no correlation between time spent and missing elements. The results are similar to reading of the text. The number of exact or similar words is not correlated with missing elements. However, the overall answer similarity correlates to the number of missing elements with similarity being lower (i.e., worse answer) when there are more missing elements. As with text, the audio is perceived as more difficult when there are more missing elements.

**Table 6.** Results for the free recall measure for the listening (audio-text) condition. One-tailed Pearson correlation significant at 0.05 level* or at 0.01 level**.

<table>
<thead>
<tr>
<th>Missing Elements According to Rule Type</th>
<th>Time Spent</th>
<th>Exact Free Recall Count</th>
<th>Semantic Free Recall Count</th>
<th>Overall Recall Similarity</th>
<th>Perceived Difficulty</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rule 1</td>
<td>.069</td>
<td>-.070</td>
<td>-.022</td>
<td>-.179*</td>
<td>.174*</td>
</tr>
<tr>
<td>Rule 2</td>
<td>.031</td>
<td>.018</td>
<td>.074</td>
<td>-.038</td>
<td>.140</td>
</tr>
</tbody>
</table>

**Discussion**

We presented the discovery and results for missing elements, and its relationship to perceived and actual difficulty of content presented as text or audio. The effect of missing elements on reader comprehension is subtle. We believe that our broad-multiple choice questions were not sufficiently sensitive to capture the impact of missing elements. However, we find that with a more sensitive measure such as free recall, there is a relationship between the number of missing elements in the content and how much was recalled by study participants. The results are similar for audio and text presentation: when the content has more missing elements less is recalled. We believe these results might be an underestimate of the impact because AMT workers tend to work fast with quick and short answers.

Furthermore, there was an interesting but complex relationship between missing elements and perceived difficulty that needs further follow up. In the audio-text condition, having more missing elements was perceived as easier: this was when the participants were listening and only had to answer a multiple-choice question. In the text-text condition, they perceived content as more difficult when there were more missing elements. Our corpus statistics study showed that missing elements are a very frequent occurrence. It is reasonable to assume that people remove functional elements for a purpose, e.g., to make the text flow better. This discrepancy between text and audio in how they are perceived gives some insight into this, but further exploration is required. Additionally, perceived difficulty cannot be ignored since it may impact how consumers process and understand health information. The Health Belief Model (HBM) (39) and the Theory of Planned Behavior (TPB) (40) have shown this impact. In a review of 24 studies, the 4th dimension in the HBM, perceived barriers, was found to be the most significant in explaining health behavior (39). Similarly, in TPB, perceived difficulty, a sub-factor for perceived behavioral control, has been found to be the stronger predictor of intentions and behavior (41). Other work has showed how perceived difficulty correlates with a decrease in the recall of information (42).

**Practical Implications**

Our work has three main practical implications. First, audio is an effective delivery medium for information. This paper represents the second study to support this. Audio as a medium for information dissemination is becoming increasingly important the growing popularity of audio for information access. Second, our combined studies show the need for evidence-based guidance on how to optimize content for text and audio. Ideally, such guidelines will be incorporated into automated or semi-automated tools to allow a broad variety of content providers to optimize their content without requiring them to have a background in linguistics or communications. Finally, in addition to features discovered in previous work, we demonstrated here a new feature and its impact on the actual and perceived difficulty of content in text and audio presentation. This study shows that it is important to specify nouns and verbs when they refer to previously presented information. This makes the content seem easier and also contributes to better recall of the information.

Future work on health literacy for patient or consumer education where there is an information exchange with laypersons utilizing audio for information delivery would benefit from evaluating content, style, and presentation for their impact on perceived and actual difficulty. Beyond these, there are also potential opportunities for examining
other dimensions that are more relevant for audio delivery, e.g., emphasis, persuasion, and bias. With audio information delivery, few evidence-based guidelines are currently available.

**Conclusion**

Our overall goal is to optimize content for delivery both via text and audio. We aim to do this by creating tools that provide evidence-based guidelines on how to change text. In a first study conducted, we found that audio is an effective medium to deliver content, an exciting and promising result particularly given the increasing prevalence of audio devices in our lives\(^{10}\). In the study presented here, we focused on one text feature, missing elements, and found that it affects both perceived difficulty and actual difficulty for both text and audio. Again, this is a promising result since it means that optimizing content will be equally effective for text and audio presentation. While the impact of missing elements may seem small, the feature appears in nearly 30% of sentences used to provide health information; even a small improvement could have a large impact.

**Acknowledgments**

Research reported in this paper was supported by the U.S. National Library of Medicine of the National Institutes of Health under Award Number 1R01LM011975 and 2R01LM011975. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

**References**

38. Leroy G, Kauchak D. A comparison of text versus audio for information comprehension with future uses for smart speakers. JAMIA Open. Accepted for Publication.
Teaching Medical Students Health Care Failure Mode and Effect Analysis: A Case Study of Inpatient Pain Management Computerized Decision Support

Blake Lesselroth, MD, MBI1, William Dudney1, Juell Homco, PhD, MPH1, Melissa Van Cain, MD, MBI1, Savanna Smith1, Audrey Corbett, MD1
1University of Oklahoma-Tulsa School of Community Medicine, Tulsa, Oklahoma, USA

Abstract

There is a pressing need to provide health professional learners experiential learning opportunities in health systems science and quality improvement. Moreover, there are several published tools to diagnose and treat health system vulnerabilities and hazards. The Health Care Failure Mode and Effect Analysis™ (HFMEA) is a systems-engineering tool that the military and aerospace industries developed to proactively identify potential errors. While this technique has been used in a range of healthcare settings, there are few reports where health professional educators have used it with learners to teach quality improvement and systems engineering methods. We describe herewith an application of HFMEA in a medical informatics professional student rotation. In this manuscript, we briefly review HFMEA theory and methods, illustrate its application to address a quality improvement initiative, and reflect upon its value – and limitations – when used in an educational context.

Introduction and Background

There is a movement in medical education to teach health systems science – sometimes referred to as the third pillar of medicine – along with the basic and clinical sciences1. Clinical workflow analysis, process redesign, and healthcare quality improvement are foundational topics in health systems science that crosswalk with applied clinical informatics2. Consequently, there is a pressing need for academic informatics departments to provide an educational program for learners to study and apply health systems science. This program should include an experiential component that enriches learning by allowing learners to apply lessons and develop new skills.

At the University of Oklahoma-University of Tulsa School of Community Medicine (OU-TU SCM), the Department of Medical Informatics offers applied informatics and data science rotations to medical residents and health professional students. These rotations include didactics, practicums, and mentored scholarship in the form of a mini-capstone addressing a health systems science topic3. The mini-capstone projects focus on enterprise-level problems. This creates opportunities to teach core informatics topics like computerized decision support (CDS), health information technology (HIT) management, interprofessional collaboration, and quality improvement methodology.

Knowing how to “diagnose” health system failures and “prescribe” implementation science solutions is an important cross-cutting competency within these educational domains. Two systems diagnosis tools useful for evaluating healthcare workflow include root cause analysis (RCA) and Health Care Failure Mode and Effect Analysis (HFMEA)™4. We teach these concepts when students embark on quality improvement or systems engineering projects. In this manuscript, we illustrate the use of HFMEA using a CDS problem, review the benefits and limitations of HFMEA, and explain how we used HFMEA to teach health systems science and quality improvement.

Problem Definition

The OU-TU SCM maintains academic relationships with several community hospitals. One of our affiliate hospitals is part of a hospital network governed by a centralized corporate authority. Corporate leaders partnered with clinical staff to create an electronic medical record (EMR) pain management bundle. The bundle includes new electronic orders and an interdisciplinary workflow. They designed the bundle with the intent to reduce inpatient opioid use and opioid related adverse events. Encouraged by initial piloting success at two network hospitals, corporate leaders approved a “big-bang” implementation across the remaining hospitals in network. At the time of our project, leadership had not yet implemented the bundle at our affiliate hospital. Anticipating a range of implementation challenges (e.g., staff education, workflow re-engineering) and unintended consequences (e.g., under or over-treating patient pain), the clinical champions consulted our medical informatics team to conduct an HFMEA. Our informatics faculty required students to participate in this project as part of the rotation practicum.
Project and Manuscript Goals and Objectives

We identified project goals and objectives based upon stakeholder expectations. Corporate quality and safety officers set a goal to improve opiate prescribing safety. Corporate management set timeline objectives for implementation. Clinical champions (i.e., our customers) set goals to limit systems-based errors, secure clinician buy-in, and increase patient satisfaction. Our medical informatics department sought to provide a prospective risk analysis and actionable recommendations. We knew at the outset that our project had to address clinicians’ goals within management’s timeline. We also needed to provide an educational experience that met course learning objectives.

The focus of this paper is to illustrate how to use the HFMEA within an informatics educational rotation to teach health systems science techniques, satisfy professional school educational program objectives, and provide consultative support to clinicians. We have several objectives with this manuscript. We intend to (1) give a brief review of HFMEA theory and methods; (2) demonstrate use of the HFMEA in a real-world situation; (3) illustrate how to integrate these methods into a professional school rotation; and (4) highlight early lessons learned and limitations. This paper should be of interest to applied informaticians, educators, and quality improvement specialists.

Theoretical Framework and Focused Literature Review

In 2001, the Department of Veterans’ Affairs (VA) adapted HFMEA from methods used by the aerospace and military sectors to identify risks in manufacturing processes. The VA method combined concepts associated with Failure Mode Effect Analysis (FMEA) and Hazard Analysis and Critical Control Point (HACCP) to proactively identify and address health system vulnerabilities. The VA also included the Safety Assessment Code (SAC) Matrix from RCA and a novel decision algorithm to prioritize corrective actions.

The HFMEA is most effective during product design, but practitioners may use it to analyze systems in a mature healthcare enterprise. It consists of five main steps: (1) selecting a process for inspection; (2) recruiting an interdisciplinary team; (3) creating a flow process map; (4) conducting a hazard analysis; and (5) formulating an action plan to address failure modes. Because HFMEA is a proactive “diagnostic tool” to predict failures, it is crucial to assemble an interdisciplinary team of subject matter experts (SMEs). This team can draw upon their collective experience when brainstorming.

![Figure 1](image-url). Sample section of an HFMEA flow process map illustrating processes, sub-process, and hazard analysis.

After identifying all failure modes, the team assigns a SAC score to each failure mode using a 16-point scale that predicts the probability and clinical severity of an event. Once scored, the team uses the HFMEA Decision Tree to prioritize action based upon hazard criticality, absence of effective control measures, and lack of detectability. The interconnectedness and complexity of healthcare systems can make a comprehensive HFMEA cost prohibitive. Therefore, one approach is to focus upon a small section of the workflow.
There are numerous studies demonstrating the value of HFMEA across a range of clinical settings including oncology, surgery, and general inpatient care\textsuperscript{6-13}. Notably, researchers have used HFMEA to diagnosis and manage health-systems pharmacy hazards leading to adverse events\textsuperscript{9,14,15}. For example, Velez-Diaz-Pallares and colleagues used HFMEA to analyze medication management on inpatient wards using computerized physician order entry and unit dose dispensing\textsuperscript{9}. They found that HFMEA helped the quality improvement team reduce inpatient prescribing errors. Anjalee and colleagues conducted a systematic review of published HFMEAs and found it to be an effective tool for reducing medication errors\textsuperscript{10}. Similarly, Faiella and colleagues concluded that HFMEAs are effective for streamlining the evaluation of complex systems, schematizing risk assessment, and selecting safety interventions\textsuperscript{6}.

**Setting**

Our affiliate hospital is an urban tertiary care facility with intensive care, surgical subspecialty services, pediatrics, and obstetrics. The clinical champions charged with implementation included a palliative care physician and two hospitalists with training in pain management and healthcare quality improvement.

The proposed pain management implementation bundle included (1) EMR order menus; (2) CDS; (3) a new interdisciplinary workflow; (3) a staff education campaign; and (4) new hospital policies. The bundle required prescribers to select therapies from a standardized list of orders for non-opioid medications, opioid medications, and non-pharmacologic pain-management alternatives. The orders direct nurses to regularly compare patient functional status to pre-defined therapeutic goals, and administer therapies in an escalating fashion when required. For example, the care team may set a therapeutic activity goal requiring a post-operative patient with a new hip arthroplasty to transfer from bed to bedside commode by the second post-op day. If the patient cannot transfer due to pain, the nurse will begin by administering non-pharmacologic therapies and non-opioid medications. If, upon reassessment, the patient fails to reach this goal, the nurse may administer oral or intravenous opioids.

Developers piloted the bundle at two hospitals and gathered preliminary data showing a reduction in total morphine equivalents (ME) administered, a decrease in the use of opioid reversal agents, and a reduction in opioid prescriptions at discharge. There was no change in patient satisfaction scores related to pain relief. Encouraged by these findings, corporate leadership authorized bundle implementation as part of a national pain management campaign. Management expected our hospital to submit for consideration any local CDS configuration requests, implement the bundle, and remove personalized clinician order sets within 60 days.

The bundle called for a seismic change in prescribing behavior and clinical workflow. Like many real-world health system implementation, the plan had several project management constraints including fixed implementation resources, an ambitious roll-out timeline, and top-down corporate messaging. Therefore, we anticipated numerous challenges and sought to identify as many failure modes as possible. We needed to prioritize failures modes as a function of risk and propose mitigating strategies that could feasibly be implemented within 30 days.

**Methods**

The OU-TU SCM Medical Informatics rotation is a two or four week rotation for medical students, physician assistant students, and residents\textsuperscript{3}. Both formats include didactics, readings, participation in departmental meetings, and a mentored practicum. Typically, the practicum requires the learners to either participate on “in-flight” projects or design a novel project with a focused research question. The rotation culminates with the students delivering a “grand rounds” style presentation on their project to staff and faculty.

For this project, we asked students to conduct an HFMEA of the multi-modal pain order set. Informatics faculty gave several lectures on HFMEA methods and furnished students with readings detailing the theory and steps for analysis. The students then gathered and reviewed data and artifacts related to the protocol. This included preliminary reports from the pilot sites, wireframes of the order menus, written specifications for decision support, and training materials. The students met with local champions to better understand the protocol, the climate of implementation, and how they could apply HFMEA to identify potential implementation barriers.

Faculty supervised students as they conducted semi-structured interviews with SMEs. Interview topics included the current-state workflow and workflow compatibility concerns, perceived usability of the new order sets, perceived usefulness of the new protocol, appropriateness for the patient population, challenges with interdisciplinary
communication, staff training needs, known implementation challenges at other sites, patient or specialty-specific implementation barriers, and the plan for patient education.

We selected a convenience sample of SMEs based upon their clinical domain knowledge and anticipated role in the future-state workflow. The sample included physicians, nurses, physical therapists, and nurse educators employed by the hospital. It was crucial to interview SMEs familiar with inpatient pain management, order entry, and bedside care. We interviewed nurse educators in the hopes of identifying training best-practices.

The students completed a modified HFMEA using the data collected from literature, semi-structured interviews, artifact analysis, and non-participant observation. Given the time constraints, we modified and simplified the HFMEA process so students could complete a preliminary analysis within the student rotation timeline. Rather than focusing on all potential failure modes, the students focused on major themes that emerged during the interviews. Using an apprenticeship model, faculty helped students identify sub-processes, failure modes, and recommendations for corrective action. The students presented their findings to department leadership and project sponsors as part of their rotation and received feedback on their work and presentation.

**Results**

*Graphically describing the process*

From the outset, the steps of our HFMEA deviated from the classic approach described by DeRosier. Typically, the project leader assembles a multidisciplinary team with SMEs and one or more advisors. The SMEs provide insight on how a process works, whereas the advisor helps the leader scope the project and complete tasks. In our case, we had a future-state workflow in hand and needed SMEs to forecast potential problems. Therefore, instead of assembling the team to map workflow, we sought out and interviewed SMEs using the process map as a guide.

We documented the future state workflow using a swim lane diagram labeled with stakeholder roles rather than using a flow process map (*Figure 3*). This is because the workflow was multidisciplinary and included branching logic, and parallel activities. We assigned numbers chronologically to processes. This created a flow map with 22 processes and 22 sub-processes. We found it helpful to cluster related steps and label according to high-level goals. Goals included: (1) initial assessment; (2) order entry; (3) order processing; (4) goal assessment; and (5) administration.

![Figure 3. Simplified flow process map of the future-state workflow. We elected to adapt a swim lane diagram to reflect the non-linear and iterative workflow that crosses multiple stakeholders and clinical settings.](image-url)
Completing the hazard analysis

In one week, the team interviewed two internal medicine residents, five acute care nurses, four maternity care nurses, and two physical therapists. Working iteratively, the faculty and students compared interview notes with the workflow diagram and conducted brainstorming sessions with local champions. The team identified 33 failure modes and four overarching themes (Table 1).

The first theme related to the clinical appropriateness of the order sets. Many clinicians believed the protocol might be inappropriate for some patients’ pain management needs. For example, patients with acute severe post-operative pain, chronic opiate prescriptions, the inability to swallow, or medication allergies may need a bespoke plan.

The second theme related to the delay between clinical assessment and medication administration. Clinicians were concerned that the elapsed time dictated by the protocol to assess analgesic effectiveness before the next medication administration would be unacceptably long. They feared this delay would negatively affect care quality and erode trust between patient and care team.

The third theme related to efficiency. Clinicians were apprehensive about the additional time required to complete tasks. Physicians believed order set complexity would increase order entry times. Nurses were equally concerned about the additional time invested in patient education, pain management counselling, and activity assessments.

The fourth theme highlighted workflow compatibility mismatches between the current-state and future-state. For example, med/surg ward nurses did not know how to reconcile discordant patient-reported pain scales with objectively observed functional performance. If a patient rated their pain 10 on a 10-point pain scale, but met a priori activity goals, should nurses administer or withhold the next analgesic dose? In a separate example, some surgical specialties, including obstetrics, use standardized pain management strategies that did not align with the new workflow. Anesthesiologists oversee patient pain management requirements following caesarean delivery and favor ketorolac, a parenteral drug used to treat moderately severe pain. The EMR order bundle did not include ketorolac.

For several reasons, the team did not use the HFMEA Hazard Scoring matrix or the HFMEA Decision Tree™. This process is time consuming and resource intensive. However, the clinical champions requested a fast turnaround. Furthermore, new clinical management policies restricted the range of potential corrective actions. Per customer request, we prioritized corrective actions based upon logistics and feasibility. We did not prioritize recommendations requiring major technology modifications, major informatics resource investments, or hospital policy revisions.

Table 1. Excerpt of our HFMEA findings including failure mode themes, failure modes, and corrective actions.

<table>
<thead>
<tr>
<th>Goal</th>
<th>Process and sub-process</th>
<th>Affected stakeholder</th>
<th>Failure mode and theme</th>
<th>Corrective action</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial</td>
<td>1. Patient needs pain medication</td>
<td>Nurse</td>
<td>1B1. Patient arrives to ward on high dose parenteral opiate - clinical appropriateness</td>
<td>Include pharmaceutical de-escalation protocol in nursing orders</td>
</tr>
<tr>
<td>assessment</td>
<td>1B. Nurse evaluates patient</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Order entry</td>
<td>8. Select the therapeutic activity goal order</td>
<td>Physician</td>
<td>8B1. The order choices are too complex with too many goals – staff efficiency</td>
<td>Include pre-set goals for a limited number of common patient use-cases</td>
</tr>
<tr>
<td></td>
<td>8B. Choose activity of daily living goal, psychosocial element goal, and mobility goal</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Corrective Actions and Recommendations

Given the project constraints and customer request, we assembled recommendations that clinical champions or local informaticians could implement within 60 days (Table 1). Our recommendations took one of the following forms: (1) order set configurations to improve usability; (2) patient communication and education materials; (3) local executive messaging; (4) patient inclusion/exclusion decision support; and (5) workflow modifications.

Placing orders for the pain management protocol entails choosing from a lengthy list of activity goals and selecting non-pharmacologic, non-narcotic, and narcotic medication orders. The prescriber must also place several corollary orders including nursing instructions and allied care consults. While the number of options affords a high degree of flexibility, this flexibility carries both learnability and complexity costs. Therefore, we proposed offering some pre-selected options to satisfy the most commonly encountered use-cases.

Nurses expressed concerns about patient reactions to the new pain protocols, hypothesizing that patients will become frustrated if new practices deviate from prior experience or fail to meet expectations. To defuse tension, improve health literacy, and direct culpability away from nurses, we suggested developing institution-branded resources for nurses to furnish to patients describing the pain management goals, program, and rationale.

The SMEs we spoke with expressed dismay over the corporate implementation strategy, arguing that the approach disenfranchised front-line clinicians. We believe the inability to participate in decision-making created a problematic climate of implementation. We recommended that local management and executive leadership conduct a series of “safety rounds” to support and reward adoption and identify and remove barriers to use.

Recognizing that some patients may not be suitable for the multi-modal pain orders, we recommended defining patient inclusion and exclusion criteria. The organization could communicate these criteria through in-person and online trainings, published materials, and point-of-care decision support. We also recommended developing alternative order pathways to accommodate patients that were not suitable for the standard orders.

Finally, we outlined several solutions to handle common workflow exceptions. For example, it may be necessary to include a protocol for patients arriving on high-dose narcotic analgesia. These patients may need a different nursing assessment strategy and an analgesic de-escalation protocol. Also, providers may need to quickly enter pain management orders for patients at hospital admission, before the inter-disciplinary team can assess the patient’s functional status. We recommended including order sets that “release” when allied team members complete the functional assessment. We also recommended using temporary pain management holding orders as a bridge until the provider can enter the multi-modal pain protocol.

Educational Impact

Students shared with faculty several valuable insights about their educational experience. The students received formal instruction and practical experience on many core informatics topics including CDS, workflow analysis, process

---

<table>
<thead>
<tr>
<th>Goal</th>
<th>Process and sub-process</th>
<th>Affected stakeholder</th>
<th>Failure mode and theme</th>
<th>Corrective action</th>
</tr>
</thead>
<tbody>
<tr>
<td>Goal assessment</td>
<td>18. Re-assess patient’s pain control</td>
<td>Nurse</td>
<td>18A1. Time elapsed between first assessment and administration of parental medication could be 2.5h – clinical delay</td>
<td>Remove third assessment step from future-state workflow</td>
</tr>
<tr>
<td></td>
<td>18A. Determine if patient is meeting current goals</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Administration</td>
<td>13. First tier pharmaceutical analgesic administration</td>
<td>Nurse</td>
<td>13B1. The patient is post-caesarian delivery and still on the anesthesia protocol – workflow compatibility mismatch</td>
<td>Identify and post patient cohorts that meet exclusion criteria</td>
</tr>
<tr>
<td></td>
<td>13B. Nurse reviews initial pharmaceutical analgesic order</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---
One weakness the students reported was the lack of prior informatics training which created a steep learning curve. They also noted that the short rotation timeline made it challenging to complete larger projects. Because most operational informatics projects continue longer than an educational rotation, it is crucial that faculty build in mechanisms to teach and support project handoffs. The lack of an established handoff format in medical informatics posed a formidable challenge.

We pragmatically adapted the situation-background-assessment-recommendations (SBAR) framework used in healthcare for patient care handoffs. The students and faculty integrated SBAR information into the final presentation. For “Situation,” students provided a concise summary of the project and relevant informatics domains. For “Background,” the students described our adaptation of the HFMEA and a brief description of the pain management protocol. For “Assessment,” the students reviewed the flow diagram, failure modes, and preliminary recommendations to the customer. For “Recommendations,” outgoing students outlined future strategies for incoming students. Students then exchanged all materials, artifacts, and data.

Discussion

Principle findings

The HFMEA is a robust systems analysis method that is ideally suited for healthcare settings where interdisciplinary teams need a structured approach to unpack, understand, and predict the behavior of complex adaptive systems. It is an effective tool for targeting workflow vulnerabilities, estimating patient safety risks, and prioritizing solutions in resource constrained settings. We found that applying the HFMEA framework to an implementation initiative offered several practical advantages. It increased awareness among stakeholders and fostered interdisciplinary engagement. It generated recommendations for order sets, educational materials, communication strategies, and special-case workflows. The HFMEA also provided a framework for stakeholder discussions and a way to track and finalize recommendations.

While healthcare organizations need competent clinical professionals applying these methods, training programs rarely teach HFMEA. We believe HFMEA is a practical and teachable method that educators can incorporate into undergraduate and graduate medical education. The informatics teaching faculty found it provided a novel and dynamic way to teach systems-based practice to health professional students. By applying HFMEA concepts to a real-world setting, students gained practical experience with quality improvement methods and adapted methods to suit the clinical and business context. Moreover, the experiential nature of the program enabled the faculty to observe and evaluate entrustable professional activities.

It is critical to point out that HFMEA can be time consuming and resource intensive. DeRosier and colleagues noted that a single HFMEA can require large interdisciplinary teams and 10 or more meetings. For this reason, they suggested only examining one facet of a process so as not to overwhelm the participants. Our project was characterized by (1) a rigid, prescriptive future-state workflow; (2) top down implementation without stakeholder input; and (3) significant time and resource constraints. Therefore, we modified the HFMEA steps and streamlined the approach to meet leadership’s deadline, identify nimble solutions, and engage novice learners. We still believe the standard HFMEA framework is an excellent method for analyzing a system; practitioners should complete each step if time and resources permit. However, healthcare executives often demand quick and decisive action. In our experience, it is important to teach students how to keep pace with business operations by adapting to the use-case and available resources.

Relevance to current literature and future steps

There are many published descriptions using HFMEA; our recent literature search using PubMed and Google Scholar, identified 131 monographs describing application in specific disciplines (e.g., radiation oncology) or processes (e.g., inpatient supply chain management). However, we found only one report describing the use of HFMEA to teach health professionals. Schuller and colleagues sent department faculty to a continuing medical education conference offered by the American Association of Physicists in Medicine (AAPM) and then gave a condensed version to department personnel in a series of lunch seminars. The seminars used a combination of didactics and use-cases.
to teach flow process mapping, failure mode analysis, and fault tree analysis. However, they did not describe the training methods, strengths, or limitations.

For several reasons, we believe HFMEA is a teachable, feasible, and generalizable systems diagnosis strategy that academic programs should include in their health systems science curriculum for health professional students, residents, and fellows. First, HFMEA is a useful technique in quality improvement work to analyze systems. Teaching this method to health professional students provides future practitioners with a practical and adaptable skill they can use to diagnosis system stress points and explore a range of solutions. Second, HFMEA provides a framework for students from different programs to leverage their unique skills and knowledge on interprofessional teams. Third, curricular modules incorporating HFMEA can be used by professional programs to meet health systems science learning objectives required by accreditation bodies.

**Strengths and limitations**

Despite the role of HFMEA in health systems science and patient safety, researchers have highlighted important methodologic limitations that could bias findings and outputs. First, identification of failure modes relies heavily upon facilitated brainstorming sessions with interdisciplinary groups. Therefore, participants are vulnerable to anchoring bias, availability bias, and group-think. For this reason, it may be useful to combine HFMEA methods with human factors research methods such as user simulations, cognitive step-throughs, or non-participant ethnography. Second, risk scoring and decision analysis have validity issues. Scoring risk based upon perceived probability and severity requires a considerable amount of guesswork. Finally, decision analysis methods in the HFMEA process do not quantify the reliability or effectiveness of system controls, backups, or fail-safe measures.

Faiella and colleagues theorized that HFMEA may miss certain classes of failure modes and recommended examining human-computer interaction errors using Systematic Human Error Reduction and Prediction Analysis (SHERPA). They also recommended analyzing the interconnectedness of complex adaptive systems using Systems Theoretic Accident Model and Processes and System Theoretic Process Analysis (STAMP-STPA). Similarly, Abrahamsen and colleagues suggested combining HFMEA with other systems engineering methods such as incident learning and Structured What If Technique (SWIFT). Kricke and colleagues recommended using EMR data and big-data analytic methods to identify sub-processes and workflow variations overlooked during process mapping.

This background literature provides insights into limitations in our work. First, the rotation schedule and management objectives created an aggressive project timeline; students had very little time to brainstorm with frontline workers. This created an inherent selection bias. We could address this risk in the future by surging resources and assigning more students to the process workflow. Second, to improve and measure the validity and completeness of HFMEAs, we need to concurrently assign several groups of learners to independently complete an HFMEA on the same system and then compare outputs. Third, we hope to add human factors methods to our analysis protocol. Fourth and finally, adding a health data science module using EMR data to identify systems issues and errors might provide an important dimension to an overall safety appraisal.

**Conclusions**

In summary, we believe that the HFMEA is an important tool for health systems diagnosis and a powerful educational lever to improve the informatics and quality improvement competencies of health professional learners. However, traditional HFMEA methods can be quite time intensive and often demand full engagement of an interdisciplinary clinical team. This can erode stakeholder enthusiasm and limit practicality. Through this use-case, we demonstrated how to adapt methods to align with the pace of business operations and offered strategies to teach HFMEA to learners.

**References**


Validation of Real-World Data-based Endpoint Measures of Cancer Treatment Outcomes

Qian Li, MS¹, Hansi Zhang, MS¹, Zhaoyi Chen, PhD¹, Yi Guo, PhD¹, Thomas J George Jr, MD, FACP¹, Yong Chen, PhD², Fei Wang, PhD³, Jiang Bian, PhD¹

¹University of Florida, Gainesville, FL, USA; ²University of Pennsylvania, Philadelphia, PA; ³Weill Cornell Medicine, New York, NY, USA

Abstract

Recently, there has been a growing interest in using real-world data (RWD) to generate real-world evidence that complements clinical trials. To quantify treatment effects, it is important to develop meaningful RWD-based endpoints. In cancer trials, two real-world endpoints are of particular interest: real-world overall survival (rwOS) and real-world time to next treatment (rwTTNT). In this work, we identified ways to calculate these real-world endpoints with structured electronic health record (EHR) data and validate these endpoints against the gold-standard measurements of these endpoints derived from linked EHR and tumor registry (TR) data. In addition, we examined and reported data quality issues, especially inconsistencies between the EHR and TR data. Using a survival model, we show that the presence of next treatment was not significantly associated with rwOS, but patients who had longer rwTTNT had longer rwOS, validating the use of rwTTNT as a real-world surrogate marker for measuring cancer endpoints.

Introduction

In clinical trials, an endpoint is a “precisely defined variable intended to reflect an outcome of interest that is statistically analyzed to address a particular research question,”¹¹ and is usually characterized by the type of research questions or outcomes the trials aim to assess. In cancer trials, the most common outcome-based endpoints are overall survival and measurements of tumor burden, such as tumor response rate and progression-free survival,²,³ that are normally used to measure how well a treatment has worked. A number of endpoints have emerged that can be used as surrogate markers for “duration of clinical benefit,”⁴,⁵ such as time to next treatment (TTNT). Although clinical trials, especially randomized clinical trials, are considered the gold standard for generating clinical evidence about a treatment and its outcomes, they are expensive, time-consuming, and require the recruitment of sufficient participants, which is often difficult⁶,⁷. Further, trial results are often not generalizable to patients treated in real-world settings due to issues such as overly restrictive eligibility criteria⁸. These issues are especially prevalent in cancer trials,⁶,⁷ and approximately 97% of oncology trials ultimately fail⁹.

Recently, there has been a growing interest in using real-world data (RWD) to generate real-world evidence (RWE) that complements the results of clinical trials. The term RWD, widely promoted by the U.S. Food and Drug Administration (FDA), refers to data collected from sources outside of conventional research settings, including electronic health records (EHRs), administrative claims, disease registries, and billing data, among others.¹⁰,¹¹. These RWD sources contain detailed, longitudinally tracked patient information such as disease status, treatment, comorbidities, and concurrent treatments. The information generated from RWD can provide valuable RWE about how patients are treated in real-world clinical settings that can inform therapeutic development, outcomes research, patient care, safety surveillance, and comparative effectiveness studies.¹² In order to quantify the treatment effects, however, it is necessary to develop meaningful RWD-based endpoints.

Another important RWD source in cancer research is tumor registry (TR) data, which is often manually extracted from cancer patients’ medical charts (i.e., EHRs). Variables related to endpoints such as overall survival (OS), cause of death, and the date when the first line of treatment was begun can be reliably obtained from TR data. However, there are a number of gaps in TR data: it lacks detailed information about patients’ other characteristics (e.g., comorbidities) and longitudinal information about patients’ cancer treatment trajectories,¹³ and it does not capture all the cancer patients in the health system, for various practical reasons (e.g., state or national TR reporting requirements and reporting delays due to the labor-intensive manual abstraction process).¹⁴. Patient data missing from the TR leads to

* Corresponding: Jiang Bian; bianjiang@ufl.edu
various issues for research studies (e.g., reduced sample size leading to reduced power of the estimates). For RWD-based research studies, therefore, using linked raw EHR and TR data is ideal.

Literature on real-world endpoints is emerging and a number of endpoints have been proposed, including real-world overall survival (rwOS), real-world time to next treatment (rwTTNT), and real-world progression-free survival (rwPFS)\(^{15-17}\). Nevertheless, the ability of researchers to use RWD to extract these endpoints remains an area of active discussion. For example, rwTTNT can be calculated based on structured EHR data alone, using the dates of various procedures and diagnoses, while rwPFS requires information on the tumor itself, which is often only available in unstructured clinical text (e.g., pathology reports).

Two real-world endpoints are particularly of interest. The first is rwOS, the time from the date of cancer diagnosis or treatment initiation (depending on the type of cancer or the study aims) to the date of death, end of follow-up, or last contact; the second is rwTTNT, the time from the initiation of the first course of cancer-directed treatment to the initiation of the next line of therapy (i.e., “subsequent treatment” in the case of recurrence or progression)\(^{15,16}\). Although rwTTNT derived from RWD sources has not been examined extensively, it can provide critical insight on the real-world performance of cancer treatments\(^{16,19}\). For example, rwTTNT can be used to estimate progression-free survival and the effectiveness of cancer treatment\(^{5}\).

In this work, we aimed to identify ways to calculate real-world endpoints with structured EHR data and validate these endpoints against the gold-standard measurements of these endpoints derived from linked EHR and TR data. We focused on two real-world endpoints, rwOS and rwTTNT, in stage I–III colon cancer patients. We used RWD from the OneFlorida Clinical Research Consortium, a Patient-Centered Outcomes Research Institute (PCORI)-funded clinical data research network contributing to the national Patient-Centered Clinical Research Network (PCORnet).

**Methods**

**Overview.** Our primary analysis goals were to assess the alignment between EHR and TR data, describe the data gaps between the two, and assess the validity of real-world endpoints derived from EHR and TR data. To achieve these objectives, we first needed to understand colon cancer patients’ cancer care pathways using RWD. A conceptual stage I–III colon cancer patient timeline is shown in Figure 1.

![Figure 1. Stage I-III Colon cancer patient treatment timeline.](image)

Our analysis involved five major parts: We (1) identified the events and associated dates of colon cancer diagnoses, death or last contact, and colon cancer treatments from both EHR and TR data; (2) explored the discrepancies between the events and event dates recorded in the EHR and TR data (e.g., patients identified as dead in the EHR but not in the TR, or vice versa); (3) computed the rwOS; (4) identified the starting points of patients’ first course of cancer treatments and subsequent treatments and computed the rwTTNT; and (5) examined the associations between the rwOS and (i) the presence of a subsequent treatment and (ii) rwTTNT. We hypothesized that patients who had a subsequent treatment would have shorter rwOS compared to those who did not and that among patients who had subsequent treatment, those who had a longer rwTTNT would have a longer rwOS than those who had a shorter rwTTNT. The typical first course of cancer treatment and subsequent treatment choices include surgery, radiation therapy, and chemotherapy. As we focused on stage I–III colon cancer patients, we only focused on systemic chemotherapy as depicted in clinical guidelines (see section below for details), and the next line of treatment refer to the intervention with a new regimen after initial chemotherapy\(^{20}\).

**Data source and study population.** We used the linked EHR and TR data from the OneFlorida network. The OneFlorida network contains linked, robust, longitudinal, patient-level RWD from approximately 15 million Floridians, including data from Medicaid claims, TR, vital statistics, and EHRs from its clinical partners. As one of the clinical research networks contributing to PCORnet, OneFlorida includes 12 healthcare organizations that provide care through 4,100 physicians, 914 clinical practices, and 22 hospitals covering all 67 Florida counties. The OneFlorida data follows the PCORnet Common Data Model (CDM) and includes patient demographics, enrollment...
status, vital signs, conditions, encounters, diagnoses, procedures, prescribing and dispensing records, lab results, etc. We extended the CDM to incorporate TR data, which follows the North American Association of Central Cancer Registries (NAACCR) standards\textsuperscript{21}. Currently, OneFlorida includes TR data from three partners that maintain records of documented neoplasms (typically malignant) in their local hospital TRs. The TR records are linked with patients’ EHRs in the OneFlorida data.

The selection of the study cohort is illustrated in Figure 2. We first identified patients diagnosed with colon cancer in both EHR or TR data. Patients were grouped into (A) patients with colon cancer diagnoses in both EHR and TR data, (B) patients with colon cancer diagnoses only in the EHR data, and (C) patients with colon cancer diagnoses only in the TR data. To ensure data quality, and because the stage information was only available discretely in the TR data, we restricted our analyses to Group A. As stage 0 patients need more information to confirm their cancer diagnoses and treatment plans for stage IV patients are more complex, we further restricted the study cohort to patients with stage I–III colon cancer. We also excluded patients with unknown or missing stage information and patients who were diagnosed with colon cancer before 2012 in TR, since our EHR data only included records dating after 2012.

**Determination of cancer cases.** To extract colon cancer patients, we used the International Classification of Disease, Ninth/Tenth Revision, Clinical Modification (ICD-9/10-CM) codes 153.* and C18.* for the EHR data and the International Classification of Disease for Oncology, 3rd Edition (ICD-O-3) codes C18.0 through C18.9 for the TR data. In the EHR data, a colon cancer patient’s onset date was defined as the earliest encounter date with the colon cancer diagnosis.

**Determination of patients’ last contact or death dates to calculate rwOS.** We calculated rwOS as the length of the period from the date of the first colon cancer diagnosis to the death or last contact date. In OneFlorida, the death records in the EHR data come from two sources: The deaths are either recorded directly in the EHR by the health system (e.g., inpatient deaths or deaths reported to the health system by relatives) or are extracted from the Social Security Administration (SSA)'s Death Master File (DMF) and third party data from public and private obituaries through a privacy-preserving record linkage process by a third-party vendor, Datavant\textsuperscript{22}. The death dates in the Datavant’s death data only contain the month and year to protect privacy; thus, we imputed the death date to the first day of the month. If a patient did not have a death record, we assumed the patient to be alive. We then used the last encounter date of the patient in the EHR system as the last contact date.

TRs typically contain a cancer patient’s vital status (i.e., alive or dead) and the death date (if dead) or the last contact date (if alive). However, in our TR data, the last contact date information was missing. To determine patients’ death or last contact date, we combined EHR and TR vital status and event dates as follows: If patients were indicated to be dead in any of the data, we treated them as dead; if they had death dates in both the EHR and TR data, we used the one from the TR data. If patients were indicated to be alive in both datasets, we treated them as alive and used the last contact date from the EHR data.

**Summarization of colon cancer treatments to determine rwTTNT.** We computed rwTTNT as the length of the period between the beginning of the first course of treatment to the beginning of subsequent treatment. To identify the colon cancer treatment, we used the EHR data, as TR data does not contain patients’ longitudinal treatment records. We reviewed the National Comprehensive Cancer Network (NCCN) Clinical Practice Guidelines in Oncology\textsuperscript{23} and found
that treatments for stage I–III colon cancer include surgery, chemotherapy, and radiation therapy. For surgery, we identified the Current Procedural Terminology (CPT) and Healthcare Common Procedure Coding System (HCPCS) codes and extracted surgery-related procedures from the EHR data. To identify chemo- and radiation therapies, we used the Cancer Therapy Look-up Tables developed by the Cancer Research Network (CRN)\textsuperscript{24}. As we focused on chemotherapy after surgery, we first identified all chemotherapy occurrences in the EHR data and further distinguished between the first course of treatment and subsequent treatments. We defined the beginning of the first course of treatment as the earliest chemotherapy record in the EHR after the patient’s colon cancer diagnosis date. We implemented two rules for identifying subsequent treatments: (1) if two adjacent chemotherapy occurrences had a gap between them of over 90 days, we considered the former occurrence as the end of the first course of treatment, and the latter occurrence as the beginning of subsequent treatment; and (2) if the patient switched to a new colon cancer treatment regimen, we considered the onset of the new regimen as the beginning of the subsequent treatment. If a patient had multiple subsequent treatments, we considered only the first occurrence in our analysis.

**Statistical analyses.** We first computed the confusion matrices for events (i.e., presence of colon cancer diagnosis and vital status) obtained from the EHR data vs. the TR data. We computed kappa coefficients for each event from the EHR and the TR data. We also compared the dates of these events between the EHR and TR data.

Further, to assess the validity of rwTTNT, we built two Cox proportional hazards models to examine the association between subsequent treatment and overall survival, one that considered whether the patients had subsequent treatments or not, and a second that considered the rwTTNT. For both models, the outcome was the rwOS determined by both the EHR and the TR data. We controlled for age at diagnosis, sex, race/ethnicity, cancer stage at diagnosis, Charlson Comorbidity Index (CCI) (based on diagnosis records from EHR data\textsuperscript{25}), and smoking status.

All data processing procedures were conducted using Python and statistical analyses were performed with SAS, version 9.4 (SAS, Cary, NC, USA).

**Results**

**Demographic characteristics.** Our final study cohort (Group A) included 1,372 stage I–III colon cancer patients. The patients’ characteristics are summarized in Table 1. The mean age at colon cancer diagnosis was 65.2 years old. Men and women were represented approximately equally (50.7% and 49.3%, respectively). The majority of the patients were non-Hispanic Whites (NHW). There were more stage III patients (42.6%) than stage II (33.7%) or stage I (23.7%). Most of the patients (81.0%) had no comorbidities defined by CCI. At baseline, 41.4% of the patients had never smoked, and 58.6% were current or previous smokers.

<table>
<thead>
<tr>
<th>Demographic Characteristic (N = 1,372)</th>
<th>Mean / N</th>
<th>SD / %</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age at diagnosis, years</strong></td>
<td>65.2</td>
<td>13.8</td>
</tr>
<tr>
<td><strong>Sex</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>676</td>
<td>49.3%</td>
</tr>
<tr>
<td>Male</td>
<td>696</td>
<td>50.7%</td>
</tr>
<tr>
<td><strong>Race/Ethnicity</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-Hispanic White</td>
<td>821</td>
<td>59.8%</td>
</tr>
<tr>
<td>Non-Hispanic Black</td>
<td>230</td>
<td>16.8%</td>
</tr>
<tr>
<td>Hispanic</td>
<td>198</td>
<td>14.4%</td>
</tr>
<tr>
<td>Other</td>
<td>123</td>
<td>9.0%</td>
</tr>
<tr>
<td><strong>Colon Cancer Stage at diagnosis</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>325</td>
<td>23.7%</td>
</tr>
<tr>
<td>II</td>
<td>462</td>
<td>33.7%</td>
</tr>
<tr>
<td>III</td>
<td>585</td>
<td>42.6%</td>
</tr>
<tr>
<td><strong>Charlson Comorbidity Index at diagnosis</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>None (0)</td>
<td>1112</td>
<td>81.0%</td>
</tr>
<tr>
<td>Mild (1–2)</td>
<td>117</td>
<td>8.5%</td>
</tr>
<tr>
<td>Moderate (3–4)</td>
<td>62</td>
<td>4.5%</td>
</tr>
<tr>
<td>Severe (≥ 5)</td>
<td>81</td>
<td>5.9%</td>
</tr>
<tr>
<td><strong>Smoking status at diagnosis</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Never</td>
<td>568</td>
<td>41.4%</td>
</tr>
<tr>
<td>Ever</td>
<td>804</td>
<td>58.6%</td>
</tr>
</tbody>
</table>
Cancer diagnosis comparison. Although all 1,372 patients had records of colon cancer diagnosis in both the EHR and the TR data, there were discrepancies in the diagnosis dates between the EHR and the TR data as shown in Table 2. The majority of the discrepancies (73.4%) were of less than 1 month. The TR diagnosis dates tended to be earlier than the EHR dates: 16.2% of patients had a TR date that was 1 to 3 months earlier than their EHR date, and 9.0% had a TR date that was more than 3 months earlier. There were very few patients whose diagnoses were recorded earlier in the EHR than the TR: 0.9% were 1 to 3 months earlier and 0.5% were more than 3 months earlier.

Table 2. Differences in colon cancer diagnosis dates between EHR and TR data.

<table>
<thead>
<tr>
<th>Colon cancer diagnosis date difference (N = 1,372)</th>
<th>Frequency</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>EHR record is 1–3 months earlier than TR record</td>
<td>12</td>
<td>0.87%</td>
</tr>
<tr>
<td>EHR record is &gt;3 months earlier than TR record</td>
<td>7</td>
<td>0.51%</td>
</tr>
<tr>
<td>TR record is 1–3 month earlier than EHR record</td>
<td>222</td>
<td>16.18%</td>
</tr>
<tr>
<td>TR record is &gt;3 months earlier than EHR record</td>
<td>124</td>
<td>9.04%</td>
</tr>
<tr>
<td>Within 1 month</td>
<td>1007</td>
<td>73.40%</td>
</tr>
</tbody>
</table>

Although our statistical analyses focused on Group A patients, we further explored Group B and Group C patients’ cancer diagnosis records, as shown in Table 3. For patients in Group B (i.e., patients who had colon cancer diagnoses only in the EHR data), more than 50% (N = 1,051) of the patients were diagnosed with rectal cancer. For patients in Group C (i.e., patients who had colon cancer diagnoses only in the TR data), 75% (N = 602) of the patients’ first cancer diagnoses were before 2012, which would not have been captured in our EHR data. Among the remaining 25% (N = 195) of patients, many of their first cancer diagnoses were secondary malignancies.

Table 3. Cancer-related diagnoses for Group B and C patients.

<table>
<thead>
<tr>
<th>Cancer-related diagnosis in TR data for Group B patients (N = 2,061)</th>
<th>Frequency (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICD-O-3 site</td>
<td></td>
</tr>
<tr>
<td>C20.* (Malignant neoplasm of rectum)</td>
<td>692 (33.58%)</td>
</tr>
<tr>
<td>C19.* (Malignant neoplasm of rectosigmoid junction)</td>
<td>359 (17.42%)</td>
</tr>
<tr>
<td>C34.* (Malignant neoplasm of bronchus and lung)</td>
<td>118 (5.88%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cancer-related diagnosis after 2012 in EHR for Group C patients (N = 195)</th>
<th>Frequency (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICD 9/10 Codes</td>
<td></td>
</tr>
<tr>
<td>C78.* (Secondary malignant neoplasm of respiratory and digestive organs)</td>
<td>36 (18.46%)</td>
</tr>
<tr>
<td>C19.* (Malignant neoplasm of rectosigmoid junction)</td>
<td>26 (13.33%)</td>
</tr>
<tr>
<td>C7A.* (Malignant neuroendocrine tumors)</td>
<td>19 (9.74%)</td>
</tr>
</tbody>
</table>

Vital status and death date comparison. In Group A, there were 248 (18.08%) patients with death information in the EHR data, while 272 (19.83%) patients were deceased according to the TR data. The confusion matrix is shown in Table 4. The Kappa coefficient was 0.66 (95% CI: 0.61–0.71). Both datasets agreed that 75.8% of patients were alive and 13.7% of patients were dead; among patients with inconsistent vital status, 4.37% were alive in the TR but deceased in the EHR and 6.12% were deceased in the TR but alive in the EHR.

Table 4. Confusion matrix for vital status in EHR and TR data.

<table>
<thead>
<tr>
<th>Frequency Percent</th>
<th>TR vital status</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Alive</td>
</tr>
<tr>
<td>EHR vital status</td>
<td></td>
</tr>
<tr>
<td>Alive</td>
<td>1040</td>
</tr>
<tr>
<td>75.8%</td>
<td>6.12%</td>
</tr>
<tr>
<td>Dead</td>
<td>60</td>
</tr>
<tr>
<td>4.37%</td>
<td>13.7%</td>
</tr>
<tr>
<td>Total</td>
<td>1100</td>
</tr>
<tr>
<td>80.17%</td>
<td>19.83%</td>
</tr>
</tbody>
</table>

We compared the death dates for those recorded as deceased in both EHR and TR data. Among those 188 patients, only one patient had a death date that differed between the TR and the EHR data by over 1 month.

Figure 3 shows the distribution of rwOS in months. There were five patients with death or last contact dates that were
earlier than their colon cancer diagnosis dates; this was likely due to data entry errors or other data quality-related issues. We removed these patients for the statistical modeling. The maximum rwOS time was 101 months and the median was 22.6 months.

Using the EHR data, we identified 174 (12.68%) patients who had subsequent treatments. Table 5 shows the presence of subsequent treatment across different cancer stages. Among stage I patients, 5.85% had subsequent treatments. For the stage II and stage III patients, the percentages were 10.82% and 17.95%, respectively. The distribution of rwTTNT is shown in Figure 4. The average time to next treatment was 9.1 months with a standard deviation of 10.6 months and the median was 6.3 months.

**Table 5. Presence of subsequent treatment by colon cancer stage.**

<table>
<thead>
<tr>
<th>Colon Cancer Stage at Diagnosis</th>
<th>Presence of Subsequent Treatment</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No (95.15%)</td>
</tr>
<tr>
<td>I</td>
<td>306</td>
</tr>
<tr>
<td>II</td>
<td>412</td>
</tr>
<tr>
<td>III</td>
<td>480</td>
</tr>
<tr>
<td>Total</td>
<td>1198</td>
</tr>
</tbody>
</table>

**Association between presence of subsequent treatment/rwTTNT and rwOS.** Table 6 shows the estimates of the Cox proportional hazards model for rwOS using the presence of subsequent treatments as a predictor. Age at diagnosis was significantly inversely associated with rwOS, with a hazard ratio (HR) of 1.038. Sex was not a significant predictor for rwOS, but the HR was 1.246, meaning males had shorter rwOS compared to females. Compared to non-Hispanic Whites, non-Hispanic Blacks and Hispanics had lower hazards, with HRs of 0.659 and 0.623, respectively. Cancer stage at diagnosis was significantly associated with rwOS: stage I and II patients had much longer rwOS compared to stage III patients, with HRs of 0.454 and 0.517, respectively. Patients with CCI > 0 had shorter rwOS compared to those with no comorbidities. Smoking status at baseline was not associated with rwOS. The presence of subsequent treatment was not statistically associated with rwOS, where the HR is 1.203 with a p-value of 0.2173.

**Table 6. Estimates of real-world overall survival with presence of subsequent treatments as a predictor.**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Hazard Ratio</th>
<th>95% Hazard Ratio Confidence Limits</th>
<th>p-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>1.038</td>
<td>1.028 - 1.048</td>
<td>&lt;.0001*</td>
</tr>
<tr>
<td>Sex</td>
<td>Male vs. Female</td>
<td>1.246</td>
<td>0.997 - 1.556</td>
</tr>
<tr>
<td>Race-ethnicity</td>
<td>0.659</td>
<td>0.473 - 0.918</td>
<td>0.0138*</td>
</tr>
<tr>
<td>Race-ethnicity</td>
<td>0.623</td>
<td>0.425 - 0.915</td>
<td>0.0157*</td>
</tr>
<tr>
<td>Race-ethnicity</td>
<td>0.911</td>
<td>0.600 - 1.383</td>
<td>0.6614</td>
</tr>
<tr>
<td>Colon cancer stage at diagnosis</td>
<td>0.454</td>
<td>0.332 - 0.619</td>
<td>&lt;.0001*</td>
</tr>
<tr>
<td>Colon cancer stage at diagnosis</td>
<td>0.517</td>
<td>0.399 - 0.670</td>
<td>&lt;.0001*</td>
</tr>
<tr>
<td>Charlson comorbidity index at diagnosis</td>
<td>1.724</td>
<td>1.343 - 2.214</td>
<td>&lt;.0001*</td>
</tr>
<tr>
<td>Smoking status at diagnosis</td>
<td>1.027</td>
<td>0.805 - 1.311</td>
<td>0.8288</td>
</tr>
<tr>
<td>Presence of subsequent treatment</td>
<td>1.203</td>
<td>0.897 - 1.614</td>
<td>0.2173</td>
</tr>
</tbody>
</table>

*: statistically significant at 0.05 level; #: close to statistically significant at 0.05 level.
Figure 5 shows the Kaplan–Meier (KM) curves for rwOS, stratified by sex, race-ethnicity, stage, CCI, and presence of next treatment. Males have lower survival probabilities than females. In Figure 5 (c), the KM curves of the presence vs. no-presence of next treatment cross each other, indicating the hazard ratio is not constant over time.

![Figure 5. Kaplan–Meier survival plots of real-world overall survival, stratified by (a) sex, (b) race-ethnicity, (c) colon cancer stage, (d) Charlson comorbidity index (CCI), and (e) presence of next treatment.](image)

Table 7 shows the estimates of the Cox proportional hazards model for rwOS with rwTTNT as a predictor on patients who had subsequent treatments. Real-world TTNT itself was statistically significant; it had an HR of 0.973, which means patients with longer rwTTNT had a longer rwOS.

**Table 7. Estimates of Cox proportional hazards model for real-world overall survival with real-world time to next treatment as a predictor.**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Hazard Ratio</th>
<th>95% Hazard Ratio Confidence Limits</th>
<th>p-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>1.005</td>
<td>0.982 1.029</td>
<td>0.6568</td>
</tr>
<tr>
<td>Sex</td>
<td>Male vs. Female</td>
<td>1.049 1.780</td>
<td>0.8599</td>
</tr>
<tr>
<td>Race-ethnicity</td>
<td>NHB vs. NHW</td>
<td>0.611 0.315 1.185</td>
<td>0.1451</td>
</tr>
<tr>
<td>Race-ethnicity</td>
<td>Hispanic vs. NHW</td>
<td>0.546 1.805</td>
<td>0.3212</td>
</tr>
<tr>
<td>Race-ethnicity</td>
<td>Other vs. NHW</td>
<td>0.791 2.289 5.625</td>
<td>0.6653</td>
</tr>
<tr>
<td>Colon cancer stage at diagnosis</td>
<td>I vs. III</td>
<td>0.362 0.199 1.096</td>
<td>0.0721*</td>
</tr>
<tr>
<td>Colon cancer stage at diagnosis</td>
<td>II vs. III</td>
<td>0.539 0.286 1.016</td>
<td>0.0559*</td>
</tr>
<tr>
<td>Charlson comorbidity index at diagnosis</td>
<td>&gt; 0 vs. = 0</td>
<td>1.165 0.633 2.142</td>
<td>0.6241</td>
</tr>
<tr>
<td>Smoking status at diagnosis</td>
<td>Ever vs. Never</td>
<td>0.746 1.517</td>
<td>0.4181</td>
</tr>
<tr>
<td>rwTTNT</td>
<td>0.973</td>
<td>0.948 0.999</td>
<td>0.0452*</td>
</tr>
</tbody>
</table>

*: statistically significant at 0.05 level; #: close to statistically significant at 0.05 level.

Figure 6 shows the KM survival plots of rwOS for the patients with subsequent treatment only, stratified by sex, race-ethnicity, stage, and CCI. Because of the small sample size, the difference between the survival curves for each stratum, except the colon cancer stage, are small.

**Discussion and Conclusion**

In this study, we used RWD from linked EHR and TR to identify and validate two real-world endpoints, rwOS and rwTTNT, in stage I–III colon cancer patients. We identified the necessary events on colon cancer patients’ treatment timelines, including date of cancer diagnosis, vital status and last contact date, and presence and dates of cancer treatments, in order to establish the two real-world endpoints. We assessed the discrepancies in these events and related dates between the EHR and the TR data. Using the longitudinal records in the EHR data, we differentiated the
first course of cancer treatment (focusing on chemotherapy) from any subsequent courses of treatment and used the associated dates to compute rwTTNT. We showed that presence of subsequent treatments was not significantly correlate with overall survival. We also showed that longer rwTTNT was significantly associated with longer rwOS.

![Image](image.png)

**Figure 6.** Kaplan–Meier survival plots of real-world overall survival for the patients with subsequent treatment only, stratified by (a) sex, (b) race-ethnicity, (c) colon cancer stage, and (d) Charlson comorbidity index (CCI).

There were discrepancies between our EHR and TR data, revealing various potential data quality issues with RWD. First, the records of colon cancer diagnosis did not always line up between the EHR and TR data; some patients had a colon cancer diagnosis recorded in only the EHR (Group B) or only the TR (Group C). Possible reasons for such discrepancies between the EHR and the TR could be misdiagnosis (e.g., patients with rectal cancer were initially misdiagnosed with colon cancer and then later determined to have rectal cancer, which was reported to the TR); reporting latency (there is typically a delay of more than 6 months in reporting to the TR because of the manual abstraction process); or patient continuity issues that lead to EHR data continuity issues (e.g., patients seeking care across different health systems). The discrepancies between the EHR data and the TR data in the dates of cancer diagnosis were, however, smaller than we expected. For those patients with consistent colon cancer diagnoses in both datasets (Group A, our final study cohort), over 70% had a difference in their diagnosis dates of less than 1 month. The differences between the EHR death dates and the TR death dates was likewise negligible, and only 1 patient had a difference in deceased dates of over 1 month. We did not anticipate that the last date of contact information for surviving patients would be missing from the TR data. Although RWD has data quality issues in completeness and accuracy, linking RWD from multiple sources such as EHRs and TRs could provide a more accurate depiction of patients’ care history and health status.

In terms of real-world endpoints, past studies like Stewart et al. 2019 have only shown a positive correlation (simply Spearman’s rank-order correlation) between rwTTNT and rwOS, without controlling for other covariates. In our study, we built two Cox proportional hazard models for survival analysis and controlled for a number of important risk factors such as age, gender, race-ethnicity, cancer stage, and CCI. In the first model, we dichotomized the rwTTNT as having vs. not having a second course of treatment, which led to a larger sample size. Thus, we had more statistical power to model and examine the relationship between rwTTNT and rwOS. However, in the Cox model, the non-significant hazard ratio of having vs. not having a second course of treatment showed no statistically significant association between presence of next treatment and rwOS. The crossing KM curves for the presence vs. no presence of subsequent treatments showed that the effects of having or not having subsequent treatments vary across time. Further in-depth studies are needed to investigate the association, such as adding interaction between presence of next treatment and cancer stage. In the second Cox model, we modeled the relationship between the actual length of the rwTTNT and patients’ overall survival and yielded a statistically significant result. As the Cox model is commonly
used in modeling clinical trial results, being able to run Cox models with real-world endpoints and RWD is significant in that RWD-based analyses can generate results compatible with clinical trials. The ability to generate valid surrogate markers, such as the rwTTNT that we investigated in this study, provides health outcomes and comparative effectiveness research investigators with a valuable new tool to leverage as large collections of RWD become increasingly available.

Our study is not without limitations. First, our sample size was relatively small, especially for the second Cox model on patients who had subsequent treatments. The cohort may not well represent the stage I-III colon cancer population. This could be a potential reason for the non-Hispanic whites having higher hazard (i.e., shorter OS) compared to non-Hispanic blacks and Hispanics. As additional sites contribute their TR data to OneFlorida, we can expand the study cohort to achieve higher statistical power. Second, cancer stage information is currently only available discretely in our TR data; however, it is also prevalent in unstructured documents stored in EHRs (e.g., pathology reports). If only EHR data are available, we can explore advanced natural language processing (NLP) tools to unlock critical information such as cancer stage and other tumor characteristics. Third, current rules for identifying the first course of treatment and subsequent treatments can be improved. For example, our rule defines a 90-day wash-out period to differentiate the end of the first course of treatment and the beginning of the subsequent treatment. We assumed that the 90-day gap could eliminate the effects of the previous treatment. However, different drugs have different wash-out periods, leading to potential misclassification of the different treatment courses. For example, some trials have defined a one-month wash-out period to eliminate the effects of the previous treatment for colon cancer. We can potentially further decompose the rules for each type of colon cancer chemotherapy. More fine-grained rules or computable phenotypes should be developed and validated for future studies.

In summary, we calculated rwTTNT and rwOS from EHR and TR data. We assessed the agreement between EHR and TR data on the colon cancer diagnosis, treatment, and survival-related events. We showed that rwTTNT is positively correlated with rwOS. Thus, rwTTNT could be an important surrogate marker for studies that utilize RWD and ultimately help to optimize colon cancer treatment to extend overall survival.

Acknowledgments

This work was supported in part by NIH grants R01CA246418, R21AG068717, and R21CA245858 and the OneFlorida Clinical Research Consortium (CDRN-1501-26692) funded by PCORI. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or PCORI.

References

Integrating Multimodal Electronic Health Records for Diagnosis Prediction

Rui Li¹, Fenglong Ma, PhD², Jing Gao, PhD³

¹Department of Computer Science and Engineering, University at Buffalo, NY, USA
²College of Information Sciences and Technology, Pennsylvania State University, PA, USA
³School of Electrical and Computer Engineering, Purdue University, IN, USA

Abstract

Diagnosis prediction aims to predict the patient’s future diagnosis based on their Electronic Health Records (EHRs). Most existing works adopt recurrent neural networks (RNNs) to model the sequential EHR data. However, they mainly utilize medical codes and ignore other useful information such as patients’ clinical features and demographics. We proposed a new model called MDP to augment the prediction performance by integrating the multimodal clinical data. MDP learns the clinical feature representation by adjusting the weights of clinical features based on a patient’s current health condition and demographics. Also, the clinical feature representation, diagnosis codes representation and the demographic embedding are integrated to perform the prediction task. Experiments on a real-world dataset demonstrate that MDP outperforms the state-of-the-art methods.

Introduction

Electronic Health Records (EHRs) are digital version of patient’s medical charts, which consist of longitudinal multimodal data including demographics, diagnosis, clinical notes and clinical features. Among predictive modeling tasks utilizing EHRs, diagnosis prediction is one of the most challenging and widely explored tasks, which aims to predict future diagnosis from the patient’s historical diagnosis record. The input to this task consists of a sequence of past patient visits with diagnosis codes (e.g., ICD9 codes) in each visit, and the output of the task should be the diagnosis codes at the next visit. The challenge of diagnosis prediction mainly exist in the following two aspects: (1) EHR data is heterogeneous and noisy. Various types of features (variables) are included in EHR, including categorical variables (e.g., medical codes), numerical variables (e.g., clinical measurements), and textual variables (e.g., clinical notes). Meanwhile, due to patients’ irregular visits and incomplete recording, there may be a lot of missing data. (2) The potential prediction space is very large. The current ICD-9-CM system consists of more than 13,000 codes, and it is difficult to make predictions given the large target space.

Recently, deep learning techniques have been widely adopted for diagnosis prediction tasks. In order to model the sequential EHR data, most approaches use the recurrent neural networks (RNNs) due to its ability of keeping track of sequential information. The key idea is to project the diagnosis codes in the sequence into low-dimensional features (i.e., embeddings) that capture the information that is the most relevant to the prediction task. This embedding learning is enabled by the RNN, and various methods differ in the strategies that are designed to learn such embeddings effectively. Among these methods, Retain and Dipole make predictions based on sequences of medical codes only, and GRAM, KAME and HAP also incorporates the hierarchical structure of the disease taxonomy. Despite these successes, existing diagnosis prediction methods still have the following limitations.

1) Most of the existing methods only use medical codes as input and ignore other useful information such as clinical features and demographics. Clinical features include vital signs and lab test results, which contain plenty of details about the patient’s symptoms and are considered as a significant complement of diagnosis codes. Similarly, patient demographics record the static information about the patient, including variables such as age, gender and ethnicity. Combining clinical features and demographics with diagnosis codes can greatly boost the prediction performance.

2) Recently, a method called MHM is developed to integrate the clinical features into the diagnosis prediction task. However, this approach simply concatenates the diagnosis codes representation and the clinical feature representation, and learns a representation for every layer of the disease taxonomy. In fact, the importance of clinical features varies enormously for different diseases. In addition, it is essential to assign different weights to clinical features based on current health condition of the patient, which can be derived based on patients’ clinical features and demographics information. This weighted mechanism helps to learn the clinical feature representation that better depicts the clinical symptoms.
(3) The granularity of time stamps when clinical features are recorded fluctuates and there may be a lot of missing data. Clinical features are recorded during a hospital stay. For different visits, the length of hospital stay varies. Meanwhile, in a hospital stay, vital signs and lab test results may not be recorded regularly, and different clinical features may be missing at different time stamps.

Motivated by these observations, we propose a multimodal diagnosis prediction model (MDP) to address these challenges. MDP takes diagnosis codes, disease taxonomy, clinical features and demographics as input, which are all relevant to the prediction task, and then feeds the input to a deep neural networks consisting of the following integral components. As shown in Figure 1, the diagnosis code encoder utilizes the disease taxonomy to learn the diagnosis code representations, and the clinical feature encoder learns the clinical feature representation by integrating the weight adjustment mechanism and the attentive clinical feature aggregation mechanism. The final representation, which combines predictive information obtained from the patient’s demographics, diagnosis codes and clinical features, is used to make the prediction. The proposed model is able to extract the relevant signals from all these valuable information sources, and effectively utilize them for challenging diagnosis prediction tasks with missing data and varied time granularity in the data.

Our main contributions are summarized as follows:

- We propose a novel and effective deep learning framework for diagnosis prediction, a very important task in health informatics. To the best of our knowledge, we are the first to integrate medical codes, clinical features and patient demographics in the diagnosis prediction task.
- We design a clinical feature weight adjustment mechanism which learns the correlation between the patient’s diagnosis codes and the clinical features, and adjusts the weight of clinical features based on the patient’s health condition and demographics.
- We incorporate an attentive clinical feature aggregation mechanism into the framework to deal with the missing data and the varying time granularity of clinical features. The aggregation mechanism assigns lower importance scores to the time stamps with missing data and captures the long-term dependencies by integrating the hidden states of clinical features at different time stamps.
- We empirically show that MDP outperforms existing methods on a real-world EHR data set. We also reveal some important properties of the proposed model that could explain the model superiority by experiments involving a quantitative analysis of clinical feature importance degrees and qualitative analysis based on case studies.

Problem Statement

**EHR Data.** For each patient, the clinical record can be viewed as a sequence of visits $V_1, \ldots, V_T$, where each visit record $V_t$ contains diagnosis information $x_t$ and clinical features $c_t$. The diagnosis information $x_t \in \{0, 1\}^{|D|}$ is a multi-hot binary vector, where $|D|$ is the number of unique diagnosis codes, and $D = \{d_1, d_2, \ldots, d_{|D|}\}$. $x_{t,i} = 1$ indicates that the patient was diagnosed with disease $d_i$ in the $t$-th visit; otherwise 0. For clinical features $c_t \in \mathbb{R}^{N \times T_t}$, $N$ is the number of clinical features that we are interested in, and $T_t$ is the length of the ICU stay in the $t$-th visit. The clinical features include vital signs and some lab test results. Besides, there is a demographic vector $p \in \{0, 1\}^r$ associated with each patient, which includes the patient’s gender, ethnicity, age and other demographic information, and $r$ denotes the number of such information.

**Disease Taxonomy.** Let $G$ denote the disease taxonomy, which contains the hierarchy of disease concepts in the form of a parent-child relationship, and the diagnosis codes in $D$ are the leaf nodes. We define $D’ = \{d_{|D|+1}, d_{|D|+2}, \ldots, d_{|D|+|D’|}\}$ as the set containing the ancestor codes, and all nodes in $G$ form the set $C = D + D’$. We construct $G$ using the multi-level diagnoses CCS categories.\[https://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp\]

**Diagnosis Prediction Task.** Based on the above notations, we define our task as follows. Given the patient’s diagnosis information $x_t$, clinical features $c_t$, demographic data $p$, and the disease ontology $G$, the goal of this task is to predict diagnosis codes of the next visit denoted as $\hat{y}_{t+1}$. \[727\]
Methods

Figure 1 shows the overview of the proposed MDP framework, which mainly contains two parts: (1) Diagnosis code encoder that utilizes the patient’s diagnosis codes and Gated Recurrent Unit (GRU) to learn the representations capturing the patient’s health conditions, and (2) clinical feature encoder which adjusts the weights of clinical features based on the current health conditions and demographic information, and finally learns the clinical feature representation during the ICU stay. Next, we describe the two parts separately and show how they can be optimized jointly.

Diagnosis Code Encoder

Diagnosis Code Embedding. In order to learn the robust embeddings of the diagnosis codes in the disease taxonomy $G$, we employ the graph embedding method GRAM. In Figure 1, leaf nodes or solid circles in $G$ represent diagnosis codes in set $D$, while non-leaf nodes or dashed circles represent more general concepts in $D'$. Every node in $G$ has a basic learnable embedding $e_i (1 \leq i \leq |D| + |D'|)$, where $|D'|$ represents the number of intermediate nodes. GRAM learns the final embedding vector of the $i$-th diagnosis code $m_i$ by combining the base embedding $e_i$ and its ancestors’ base embeddings via the graph-based attention mechanism. After concatenating the diagnosis embedding vectors $m_1, m_2, \ldots, m_{|D|}$ of all diagnosis codes, we obtain the diagnosis code embedding matrix $M \in \mathbb{R}^{d_1 \times |D|}$, where $d_1$ is the dimension size of the embedding vector. For the $t$-th visit, given the diagnosis information $x_t$, the vector $v_t$ is computed as $v_t = \text{tanh}(Mx_t)$.

Visit Representation Learning. After obtaining the vector $v_t$ that contains all diagnosis information at the $t$-th visit, we use Recurrent Neural Networks (RNNs) to capture the dependencies among multiple visits. RNN is an efficient method to model sequential data, and it has been widely used on healthcare data [10]. RNNs come in multiple variants, including Long-Short Term Memory (LSTM) and Gated Recurrent Unit (GRU). In this paper, we choose GRU. The hidden state $h_t \in \mathbb{R}^{g_1}$ is calculated recurrently by $h_t = GRU(v_t, h_{t-1}, \Theta_h)$, where $g_1$ is the dimension size of the hidden state, $v_t$ is the diagnosis codes vector, $h_{t-1}$ is the hidden state of the previous visit, and $\Theta_h$ represents all GRU parameters to be learned. We use $h_t$ to denote the representation of the $t$-th visit.

Clinical Feature Encoder

Clinical features include vital signs and lab test results, and they are considered as significant complements of diagnosis codes. During the ICU stay in the $t$-th visit, the vital signs such as heart rate and blood pressure are recorded hourly, and the lab tests such as blood chemistry analysis and urine analysis are performed periodically. We represent the clinical features as $c_t \in \mathbb{R}^{N \times T_t}$, which have two characteristics. (1) The importance of clinical features varies enormously for different diseases. For example, for patients with diabetes, glucose is more important comparing with body temperature, while for patients with fever, the opposite happens. (2) The timescale of clinical features fluctuates.
For different visits, the length of the ICU stay fluctuates from less than 24 hours to more than 500 hours. For visits with long ICU stay, RNNs may fail to capture the long term dependency. To take these unique characteristics into consideration, we design two corresponding modules in MDP, which are clinical feature weight adjustment and attentive clinical feature aggregation.

**Clinical Feature Weight Adjustment.** Given the patient’s current health condition \( h_t \) and demographic data \( p \), the weight adjustment system adjusts the importance of clinical features dynamically, maintaining those features that are highly relevant to the diagnosis codes, and compressing others that are less relevant. Next, we provide the details of adjusting clinical feature weights.

Given the clinical features \( c_t \), we learn the embedding \( c'_t \in \mathbb{R}^{d_2 \times T} \) with \( c'_t = W_c c_t + b_c \), where \( d_2 \) is the dimension size, \( W_c \in \mathbb{R}^{d_2 \times N} \) and \( b_c \in \mathbb{R}^{d_2} \) are the parameters to be learned. The diagnosis codes representation for every time stamp \( t \) contains the clinical information for the \( i \)-th time stamp in the \( t \)-th visit. The clinical embedding \( c'_{t,i} (1 \leq i \leq T_t) \) is fed into the GRU denoted as clinical \( GRU_\alpha \). The hidden state \( g_{t,i} \in \mathbb{R}^{g_2} \) is calculated recurrently by \( g_{t,i} = GRU_\alpha (c'_{t,i}; g_{t,i-1}; \Theta_\alpha) \), where \( g_2 \) is the dimension size, and \( \Theta_\alpha \) is the corresponding parameter.

In order to adjust the importance of clinical features, we need to learn the weights of all clinical features for every time stamp by considering both current health conditions and demographic data. In particular, we first learn a vector \( k \) that encodes the current health condition and demographic data. Given the demographic information \( p \in \{0, 1\}^r \), the embedding \( p' \in \mathbb{R}^{r_1} \) is calculated by \( p' = W_p p + b_p \), where \( r_1 \) is the dimension size, \( W_p \in \mathbb{R}^{r_1 \times r} \) and \( b_p \in \mathbb{R}^{r_1} \) are the parameters to be learned. The diagnosis codes representation \( h_t \) contains the health condition. Then we concatenate \( h_t \) and \( p' \), and compute \( k = W_o [h_t \oplus p'] \), where \( g_2 \) is the dimension size and \( W_o \in \mathbb{R}^{g_2 \times (r_1 + g_1)} \) is the parameter to be learned.

For every time stamp \( t \), we learn a vector \( \alpha_{t,i} \in \mathbb{R}^{g_2} \) with \( \alpha_{t,i} = tanh(k \odot g_{t,i}) \), where \( \odot \) is the Hadamard product. \( \alpha_{t,i} \) contains the correlation between the clinical features and the patient’s health condition at the \( i \)-th visit. Then we map the correlation vector \( \alpha_{t,i} \in \mathbb{R}^{g_2} \) into the clinical embedding space via \( \alpha'_{t,i} = tanh(W_\alpha \alpha_{t,i} + b_\alpha) \), where \( W_\alpha \in \mathbb{R}^{d_2 \times g_2} \) and \( b_\alpha \in \mathbb{R}^{d_2} \) are the parameters to be learned, and \( \alpha'_{t,i} \in \mathbb{R}^{d_2} \). Here \( \alpha'_{t,i} \) is regarded as the weights of clinical features at the \( i \)-th time stamp. We compute the weighted clinical embedding \( c''_{t,i} \in \mathbb{R}^{d_2} \) by multiplying the original clinical embedding \( c'_{t,i} \) by its corresponding weights, which is \( c''_{t,i} = \alpha'_{t,i} \odot c'_{t,i} \). In such a way, \( c''_{t,i} \) assigns more weights to clinical features that are highly related with the patient’s current health condition.

**Attentive Clinical Feature Aggregation.** After we obtain the weighted clinical embedding \( c''_{t,i} (1 \leq i \leq T_t) \), we use another GRU, clinical \( GRU_\beta \), to capture the dependencies among the multiple time stamps. The hidden state \( q_{t,i} \in \mathbb{R}^{g_3} \) is computed with \( q_{t,i} = GRU_\beta (c''_{t,i}; q_{t,i-1}; \Theta_\beta) \), where \( g_3 \) is the dimension size, \( \Theta_\beta \) is the parameter to be learned.

Because some length of the ICU stay may be extremely long, GRU may fail to learn long-range dependencies. Instead of using the hidden state at the last time stamp \( q_{t,T_t} \) as the clinical feature representation, we import the attentive clinical feature aggregation mechanism which combines the hidden states among all time stamps. For time stamp \( i, 1 \leq i \leq T_t \), the aggregation mechanism computes the corresponding attention weight \( \beta'_i \), which is a scalar, and then calculates the weighted sum of the hidden states as the final clinical feature representation \( q'_i \in \mathbb{R}^{g_3} \). The attention weight \( \beta'_i \) is computed as follows:

\[
\beta_i = w_\beta q_{t,i} + b_\beta, \quad \beta'_i, \beta'_2, \cdots, \beta'_{T_t} = softmax(\beta_1, \beta_2, \cdots, \beta_{T_t})
\]

where \( w_\beta \in \mathbb{R}^{g_3} \) and \( b_\beta \in \mathbb{R} \) are the parameters to be learned, \( \beta'_i \) and \( \beta_i \) are scalars. Finally, the clinical feature representation \( q_i \) is calculated by \( q'_i = \sum_{i=1}^{T_t} \beta'_i q_{t,i} \).

**Joint Optimization.**

We concatenate the diagnosis codes representation \( h_t \), the demographic embedding \( p' \), and the clinical feature representation \( q_i \) to obtain the vector \( s_t \) that encodes the overall patient’s health status \( s_t = [h_t \oplus p' \oplus q_i] \), \( s_t \) is then fed into a softmax layer to predict the diagnosis codes in the next visit, which is denoted as \( \hat{y}_{t+1} = Softmax(W_s s_t + b_s) \), where \( W_s \in \mathbb{R}^{P \times (g_1 + r_1 + r_2 + g_3)} \) and \( b_s \in \mathbb{R}^{P} \) are the parameters to be learned.
Table 1: Statistics of dataset

<table>
<thead>
<tr>
<th>Statistics</th>
<th>MIMIC-III</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of patients</td>
<td>5,033</td>
</tr>
<tr>
<td>Number of visits</td>
<td>13,096</td>
</tr>
<tr>
<td>Average number of visits per patient</td>
<td>2.60</td>
</tr>
<tr>
<td>Number of unique diagnosis codes</td>
<td>4,093</td>
</tr>
<tr>
<td>Average number of diagnosis codes per visit</td>
<td>13.10</td>
</tr>
<tr>
<td>Maximum number of diagnosis codes per visit</td>
<td>39</td>
</tr>
<tr>
<td>Number of unique CCS group codes</td>
<td>476</td>
</tr>
<tr>
<td>Average number of CCS group codes per visit</td>
<td>11.59</td>
</tr>
<tr>
<td>Maximum number of CCS group codes per visit</td>
<td>34</td>
</tr>
<tr>
<td>Number of clinical features</td>
<td>17</td>
</tr>
<tr>
<td>Average number of hours per visit</td>
<td>164.40</td>
</tr>
<tr>
<td>Maximum number of hours per visit</td>
<td>500</td>
</tr>
</tbody>
</table>

We compute the cross entropy loss between the ground truth $y_t$ and the predicted $\hat{y}_t$ using

$$
\mathcal{L} = -\frac{1}{T-1} \sum_{t=1}^{T-1} (y_t \log(\hat{y}_t) + (1 - y_t) \log(1 - \hat{y}_t)).
$$

This loss is calculated for a certain patient, and we compute the loss of all patients by averaging $\mathcal{L}$.

Experiments

In this section, we first introduce the experimental settings and then demonstrate the performance of the proposed algorithm on a public EHR dataset MIMIC-III. Moreover, we analyze the importance of clinical feature and visualize the clinical feature importance of three patients with different diseases. Finally, a case study is conducted to illustrate that MDP is able to correctly predict more diagnosis codes comparing with baselines.

Experimental Settings

**Dataset.** MIMIC-III is a publicly available EHR dataset, which consists the admission records of ICU patients over 11 years. EHR data include diagnosis information such as diagnosis codes and procedure codes, clinical features such as vital signs and lab tests results, and clinical notes charted by care providers. We select the patients who made at least two visits. Following the previous work, we only select patients that are older than 18 and have single ICU stay per admission. Table 1 shows the details about the dataset. When performing the diagnosis prediction task, instead of predicting diagnosis categories like most of the previous research, we aim to predict the real diagnosis codes. This means that our task is more difficult since the target space is much larger.

**Baselines.** We select seven baselines that can be divided into three groups. Group 1 contains models that do not use the disease taxonomy information, which includes RNN and Dipole. Group 2 contains models that use the disease taxonomy information, and these models only use patients’ historical diagnosis codes as input, including GRAM, KAME, and HAI. Models in Group 3 use the disease taxonomy information and use the multimodal data as input, including CAMT and MHM.

**Implementation Details.** Following previous work, we extract 17 clinical features, including heart rate, mean blood pressure, glucose and other vital signs. For every clinical feature, we assign a binary variable indicating whether the clinical feature was observed at the current time stamp, and we fill in the missing clinical features with the values in the previous time stamp. For ICU stays with long length, we only use the records in the first 500 hours. The dimension size $N$ of the clinical feature is set to 76. The demographic data is preprocessed as the method used in CAMT, and the dimension size $r$ is 11, which contains 2 genders, 5 age groups and 4 admission types. The clinical embedding size $d_2$ is set to 76, and $d_1$, $g_1$, $g_2$ and $g_3$ are set to 128.
with learning rate 0.001 and weight decay 0.001. For all models, the dimension of the diagnosis code embedding is set to 128, and the corresponding hidden state size $g_1$ is set to 128. We run every experiment ten times, and the average values and standard deviations are reported.

**Evaluation Metric.** Following CAMI[11], we use Recall@$K$ and MAP@$K$ as the evaluation criteria. For every visit $V_t$, we get a 1 if the target diagnosis code appears in the top $k$ predictions and 0 otherwise. Recall@$K$ is defined as the number of diagnosis codes that are predicted correctly in the top $k$ of $\hat{y}_t$ divided by the total number of diagnosis codes in $V_t$. MAP@$K$ refers to mean average precision, and it considers not only the precision and accuracy but also the order of diagnosis codes which are predicted correctly. We vary $K$ from 20 to 60.

**Performance Comparison**

Table 2 shows the performance of MDP comparing with seven baselines. The proposed model MDP outperforms all baselines and achieves 1.2% higher Recall@$K$ and MAP@$K$ over the best baseline. This demonstrates the effectiveness of the clinical features in diagnosis prediction task. Meanwhile, MDP is able to adjust the weights of clinical features according to the patient’s current health condition and the demographics. Among these baselines, methods in Group 1 do not use the disease taxonomy information, and they directly learn the diagnosis code embedding from the input data. The performance of RNN is worse than that of Dipole. It is because Dipole uses the bidirectional RNNs and apply the location-based attention mechanism when it makes predictions, which are able to capture visit dependency even for long sequences.

Three baselines in Group 2 all use the disease taxonomy information, and the inputs are the patient’s diagnosis codes. The performance increases from GRAM to HAP. GRAM only updates the embedding of the leave nodes, and uses the diagnosis code representation to make predictions. KAME is built upon GRAM but learns a knowledge vector that contains the coarse-grained information of ancestor codes. KAME outperforms GRAM, and it suggests that general knowledge of the ancestors helps to represent the patient’s health condition and further boosts the performance of the diagnosis prediction. However, both GRAM and KAME ignore the order among the ancestors. For example, in Figure 1 node $c_0$, $c_1$ and $c_2$ are the ancestors of $c_3$. When learning the diagnosis code embedding, GRAM and KAME treat $c_0$, $c_1$ and $c_2$ equally, without considering their order. HAP fills the gap by designing a two-round attention propagation mechanism, and the embedding of diagnosis code is updated layer by layer. The performance of HAP exceeds GRAM and KAME, which indicates that using the full ontology hierarchy improves the models’ expressibility.

Methods in Group 3 uses the multimodal data as input, and the performance of CAMP and MHM exceeds HAP. CAMP imports a memory network to save the fine-grained patient conditions, and the patient’s demographic data cooperate with the READ/WRITE operations of the memory network. Instead of using the disease hierarchy to learn a more robust embedding, MHM uses the ontology hierarchy in a different way. It models the diagnosis prediction

<table>
<thead>
<tr>
<th>Method</th>
<th>Recall@$K$</th>
<th>MAP@$K$</th>
</tr>
</thead>
<tbody>
<tr>
<td>K = 20</td>
<td>K = 40</td>
<td>K = 60</td>
</tr>
<tr>
<td>RNN</td>
<td>0.350 ± 0.004</td>
<td>0.459 ± 0.005</td>
</tr>
<tr>
<td>Dipole</td>
<td>0.359 ± 0.003</td>
<td>0.474 ± 0.002</td>
</tr>
<tr>
<td>GRAM</td>
<td>0.355 ± 0.002</td>
<td>0.474 ± 0.002</td>
</tr>
<tr>
<td>KAME</td>
<td>0.363 ± 0.003</td>
<td>0.481 ± 0.003</td>
</tr>
<tr>
<td>HAP</td>
<td>0.370 ± 0.003</td>
<td>0.485 ± 0.003</td>
</tr>
<tr>
<td>CAMP</td>
<td>0.372 ± 0.002</td>
<td>0.490 ± 0.003</td>
</tr>
<tr>
<td>MHM</td>
<td>0.371 ± 0.003</td>
<td>0.487 ± 0.003</td>
</tr>
<tr>
<td>MDP</td>
<td>0.383 ± 0.003</td>
<td>0.501 ± 0.002</td>
</tr>
</tbody>
</table>

We randomly split the datasets into the training, validation and testing sets based on the number of patients in a 0.75:0.1:0.15 ratio. For fair comparison, all models are implemented with Pytorch, and we use the Adam optimizer.
task as a hierarchical multi-label classification problem, and clinical features and diagnosis codes are used as inputs. The performance of CAMP is better than MHM because the hierarchical multi-label model generated by MHM can not fully use the ontology hierarchy. The proposed model MDP uses the patient’s diagnosis code, clinical features and demographic data as inputs, and the weights of clinical features are adjusted by the patient health condition and demographic data dynamically. The clinical feature representation learned by MDP is used as a complement of diagnosis code representation, and it contains plenty of details about the patient’s symptoms. Comparing with the best baseline CAMP, MDP gets 1% mean improvement of Recall@K and 1.2% mean improvement of MAP@K.

Clinical Feature Importance Analysis

To demonstrate the benefit of clinical feature weight adjustment and attentive clinical feature aggregation mechanism in diagnosis prediction task, we define a criteria to evaluate the importance of clinical features for every time stamp. For the \( i \)-th time stamp in the \( t \)-th visit, the clinical feature importance \( \omega_i \in \mathbb{R}^N \) is calculated by \( \omega_i = \beta_i \alpha'_{t,i} \), where \( \alpha'_{t,i} \in \mathbb{R}^{d_2} \) is a vector learned in the clinical feature weight adjustment section, and \( \beta_i \) is a scalar learned in attentive clinical feature aggregation section. Since \( \omega_i \) may contain negative numbers, we use the abs function to compute the absolute value, i.e., \( \omega_i = \text{abs}(\beta_i \alpha'_{t,i}) \). Meanwhile, categorical features are encoded into multiple dimensions. In order to fairly evaluate the contributions of all features, we compute the average as the final importance. We concatenate \( \omega_i \) for \( 1 \leq i \leq T_t \) and obtain the importance matrix of clinical features \( \omega \in \mathbb{R}^{T_t \times N} \).

**Table 3:** Diagnosis Codes for Patients in Weight Analysis

<table>
<thead>
<tr>
<th>Patient ID</th>
<th>Diagnosis Codes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Symptoms involving respiratory system and other lung symptoms(518.81, 482.42, 714.81, 519.3), Anemia(285.9), Hypothyroidism(244.9), Hypotension(458.9)</td>
</tr>
<tr>
<td>2</td>
<td>Symptoms involving heart disease(414.01, V45.82, V45.81, E87.90, 410.31), Hypertension(401.9) Symptoms involving cerebrovascular disease(438.6, 438.20), Diabetes(250.00)</td>
</tr>
</tbody>
</table>

Figure 2 shows the heatmap of the clinical feature importance matrix for two patients in a visit. In Figure 2, the \( x \) axis is the clinical feature, and the \( y \) axis is the time stamp. The color represents the feature importance, the darker color means the more feature importance. We first analyze the importance of different clinical features. Table 3 lists the corresponding diagnosis codes in the visit. We can observe that patient 1 mainly suffered respiratory diseases, and the heatmap shows that clinical features related with respiratory system obtain more weights. For example, fraction
inspired oxygen, oxygen saturation and respiratory rate have greater weights. Meanwhile, the patient also suffered anemia and hypothyroidism, and both diseases can result in hypoglycemia. We can observe that glucose also has higher weight. Patient 2 suffered heart diseases and cerebrovascular diseases, and MDP assigns high weights to respiratory rate and heart rate. Besides, this patient also suffered diabetes, and glucose is also highly weighted. The above observation shows that clinical feature weight adjustment mechanism can adjust the weights of clinical features based on the patient’s health condition.

We then analyze the importance of different time stamps. We observe that the importance increases as time goes by. In general, MDP assigns more weights to the recent time stamps. It is because clinical features in the latter time stamps can better represent the patient’s symptoms in the current visit, and these time stamps are more helpful to predict the diagnosis codes in the next visit. For patient 2, intermediate time stamps are assigned lower weights comparing with the before and after time stamps. This is because there are lots of missing clinical features in these intermediate time stamps. The above observation shows that the attentive clinical feature aggregation mechanism can assign different weights to different time stamps.

In conclusion, the heatmap of clinical feature importance illustrates that both clinical feature weight adjustment mechanism and the attentive clinical feature aggregation mechanism can coordinate to adjust the importance of clinical features and time stamps, which further helps to learn the clinical feature representations.

### Case Study

Table 4 shows the correctly predicted diagnosis codes of another two patients when $K = 20$. For patient 1, the ground truth contains 14 diagnosis codes. The proposed model MDP correctly predicts 7 diagnosis codes in the top 20 predictions. Dipole achieves the same performance, while other baselines correctly predicted 4-5 diagnosis codes. We can observe that patient 1 mainly suffered from heart diseases, MDP correctly predicts not only the diagnosis codes directly related to heart disease but also the ones describing the details, such as V45.82, which is percutaneous transluminal coronary angioplasty status. This illustrates that the clinical feature can provide more details about the disease and is helpful to make the prediction. Similarly, patient 2 has 5 diagnosis codes, and MDP correctly predicts 3 diagnosis codes in the top 20 predictions, while most baselines only predict 1-2 diagnosis codes. These results show the superiority of the proposed MDP.

<table>
<thead>
<tr>
<th>Method</th>
<th>Patient 1</th>
<th>Patient 2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Ground Truth</strong></td>
<td>250.63, 414.01, V45.82, 536.3, 585.6, 790.7, 285.21, 428.0, 041.19, 337.1, 428.22, 403.01, 999.31, 414.11</td>
<td>276.8, 518.81, 070.70, 291.81, 070.30</td>
</tr>
<tr>
<td>RNN</td>
<td>428.0, 285.21, 585.6</td>
<td>518.81</td>
</tr>
<tr>
<td>Dipole</td>
<td>536.3, 428.0, 403.01, 585.6, 414.01, 250.63, 285.21</td>
<td>291.81, 070.70</td>
</tr>
<tr>
<td>GRAM</td>
<td>585.6, 285.21, 403.01, 536.3</td>
<td>291.81, 070.70</td>
</tr>
<tr>
<td>KAME</td>
<td>585.6, 536.3, 414.01, 250.63, 285.21</td>
<td>518.81, 291.81</td>
</tr>
<tr>
<td>HAP</td>
<td>536.3, 428.0, 250.63, 585.6, 414.01</td>
<td>070.70, 291.81, 518.81</td>
</tr>
<tr>
<td>CAMP</td>
<td>585.6, 285.21, 536.3, 428.0, 250.63</td>
<td>070.70, 291.81</td>
</tr>
<tr>
<td>MHM</td>
<td>414.01, 536.3, 585.6, 403.01</td>
<td>070.70, 518.81</td>
</tr>
<tr>
<td><strong>MDP</strong></td>
<td>536.3, 414.01, V45.82, 428.0, 250.63, 585.6, 285.21</td>
<td>291.81, 518.81, 070.70</td>
</tr>
</tbody>
</table>

### Related Work

Diagnosis prediction, which aims to predict the patient’s future health condition based on their historical EHRs, is an important task in health informatics, and thus has been widely studied. Most of the previous studies investigate how to effectively utilize the patient’s historical diagnosis codes for the prediction. Some representative methods are...
discussed below. RETAIN applies an RNN with reverse time ordered EHR sequences, and designs a two-level neural attention model to provide detailed interpretation of the prediction results. Dipole designs three different attention mechanisms to learn the diagnosis code representations, and feeds the representations into a bidirectional GRU. The hidden state of the GRU is used to predict the potential diagnosis codes in the next visit. Note that diagnosis codes and medical concepts are naturally organized in a hierarchy, and the ancestor node information plays an important role in the prediction of the diagnosis code corresponding to the descendant node. Therefore, several studies exploit this hierarchical structure in diagnosis prediction. GRAM incorporates a graph-based attention mechanism, and computes the embedding of the diagnosis code as the weighted sum of the basic embedding of itself and its ancestors. KAME learns the ancestor representation and concatenates it with the diagnosis code representation. The ancestor representation contains the coarse-grained health condition. HAP proposes a hierarchical attention mechanism in which the attention propagates across the entire hierarchy from layer to layer. The embedding of the node absorbs knowledge from not only its ancestors, but also its descendants, siblings and even some distant nodes. Different from the aforementioned methods that rely on visit sequences of diagnosis codes and their relations only, the proposed MDP model incorporates additional valuable information including patient demographics and clinical features to boost the performance of diagnosis prediction.

Recently, researchers also explore the use of multimodal data to perform the diagnosis prediction task. CAMT incorporates patients’ demographics information in the prediction model. Specifically, it imports an external memory network which saves the fine-grained health condition for every top level categories in the disease taxonomy, and combines the diagnosis codes information with the patient’s demographics to help the READ/WRITE operation of the memory network. Clinical features, which also encodes predictive information for diagnosis prediction and serves as an input to the proposed MDP model, are not considered by the CAMP model. MHM uses the diagnosis codes and clinical features as input. MHM models the diagnosis prediction task as a hierarchical multi-label classification problem, and the disease taxonomy is used to generate the hierarchical label. MHM learns a representation for every layer of the disease taxonomy, and the global representation is the weighted sum of the layer representation. However, MHM neglects the relation between the diagnosis codes and clinical features and ignores the variance in patient groups defined by demographics. The proposed model MDP overcomes this shortcoming by importing a clinical feature weight adjustment mechanism. By this mechanism, the importance of clinical features is adjusted according to the health condition and the demographics.

The proposed MDP and the aforementioned models all aim to tackle the diagnosis prediction task. Existing models utilize one or two sources of information, while the proposed model takes advantage of all the relevant information including diagnosis code sequences, hierarchy, clinical features and patient demographics. Another relevant research topic is the analysis of clinical features, but note that the task tackled by these methods is not diagnosis prediction. Concare learns the inter-dependencies among clinical features, and it improves the multi-head self-attention via the cross-head decorrelation.

Conclusions

In this paper, we propose a novel model MDP to perform the diagnosis prediction task. MDP takes the diagnosis codes, clinical features and demographics as input and consists of integral components that effectively integrates these heterogeneous information sources. The diagnosis code encoder in MDP utilizes the disease taxonomy to learn the diagnosis code representations which capture the patient’s current health condition. The clinical feature encoder learns the clinical feature representations by incorporating a weight adjustment mechanism and an attentive clinical feature aggregation mechanism. The weight adjustment mechanism adjusts the weights of clinical features based on the diagnosis information and demographics, and the attentive clinical feature aggregation mechanism enables the capture of the long term dependencies among the patient’s conditions at multiple time stamps during an ICU stay. By integrating these components, the proposed MDP is able to extract meaningful signals that are relevant to diagnosis prediction from any available source. Experimental results on a real-world EHR dataset show the effectiveness of MDP for diagnosis prediction. To analyze the insights behind the proposed framework, we further demonstrate the assigned weights to various clinical features and time stamps via the clinical feature encoder as well as a comparison with baselines based on the correctly predicted diagnosis codes in the top 20 predictions for some case studies. These results explain how the proposed MDP achieves the superior performance.
Acknowledgements
This work is sponsored by NSF-IIS 1553411. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References
Testing of a Risk-Standardized Major Bleeding and Venous Thromboembolism Electronic Clinical Quality Measure for Elective Total Hip and/or Knee Arthroplasties

Troy Li, BS1, Mica Curtin-Bowen, BA1, Avery Pullman, BS1, Stuart Lipsitz, ScD1,2, Ania Syrowatka, PhD1, Michael Sainlair, MS1, Tien Thai, BS1, Alexandra Businger MPH1, Aileen Davis, PhD1, Jay R. Lieberman, MD1, Bonnie Blanchfield, ScD1, 5, David W. Bates, MD1,2, Patricia C. Dykes PhD1,2

1Brigham and Women’s Hospital, Boston, MA; 2Harvard Medical School, Boston, MA; 3University of Toronto, Ontario, CA; 4Keck School of Medicine of USC, Los Angeles, CA; 5Harvard TH Chan School of Public Health, Boston, MA

Abstract:
Brigham and Women’s Hospital has received funding from the Centers for Medicare and Medicaid Services to develop a novel electronic clinical quality measure to assess the risk-standardized major bleeding and venous thromboembolism (VTE) rate following elective total hip and/or knee arthroplasty. There are currently no existing measures that evaluate both the bleeding and VTE events following joint arthroplasty (TJA). Our novel composite measure was tested within two academic health systems with 17 clinician groups meeting the inclusion criteria.

Following risk adjustment, the overall adjusted bleeding rate was 3.87% and ranged between 1.99% – 5.66%. The unadjusted VTE rate was 0.39% and ranged between 0% – 2.65%. The overall VTE/Bleeding composite score was 2.15 and ranged between 1.15 – 3.19. This measure seeks to provide clinician groups with a tool to assess their patient bleeding and VTE rates and compare them to their peers, ultimately providing an evidence-based quality metric assessing orthopedic practices.

Introduction
Total hip arthroplasty (THA) and total knee arthroplasty (TKA) represent the most common implant surgeries performed on Medicare beneficiaries1 and are increasingly performed for younger and more active recipients2. In 2018, over 650,000 primary TKA and 374,000 THA surgeries were performed in the United States with a projected volume increase of over 100% in the next decade3–5. Patients undergoing these procedures are at risk of developing a venous thromboembolism (VTE), which includes deep vein thrombosis and pulmonary embolism. Studies have estimated that about 4.7% of patients undergoing total joint arthroplasty (TJA) would have symptomatic VTE without prophylaxis6 and that 5-14% of patients diagnosed with a VTE following TJAs are re-admitted to the hospital7. Furthermore, VTE events following TJA are potentially life-threatening and can result in chronic complications with generally poor prognoses, such as thromboembolic pulmonary hypertension8–10. Therefore, there is general agreement that VTE prophylaxis is necessary following TJA11–13. However, orthopedic surgeons face the challenge of balancing VTE prophylaxis against their associated risks which include bleeding, prosthetic joint infections, the need for reoperation, and readmission14,15. A study by Stokes et al indicated that for hip and knee surgeries, bleeding complications increased the average cost and length of stay by 30%16. Thus, it is critical for orthopedic surgeons to strike the proper balance between efficacy and safety when choosing a VTE prophylactic regimen, especially for patients undergoing TJA.

At the time of measure conception, our objective was to develop a novel eCQM (electronic clinical quality measure) to evaluate the rate of major bleeding events following elective primary THA and TKA. Acknowledging the dangers of VTE events on the opposing end of the anticoagulant use spectrum, this measure was later expanded to evaluate both VTE and major bleeding events and seeks to assess the balance between over-prescribing (increasing the risk of bleeding) and under-prescribing anticoagulants (increasing the risk of VTE) following surgery. This eCQM will provide a data driven approach to assist providers in finding the appropriate VTE prophylactic regimen that accounts for both general and patient-specific risks of VTE and major bleeding. Developing an eCQM that relies on routinely captured EHR data ensures that the clinician reporting burden is not increased. This risk adjusted measure seeks to provide clinicians with a unique tool to assess their own performance and compare it to their peers, ultimately encouraging evidence-based quality improvement.
Methods

Patient Population and Data Sources:
Data were collected and analyzed from two geographically distant academic health systems. Site 1 used the EHR vendor Epic, provided data from 2016-2019, and included six clinician groups. Site 2 used the EHR vendor Cerner, provided data from 2017-2019, and included eleven clinician groups. All eligible patients (specifications below) were included in the analysis. Information documented in the EHR (e.g., comorbidities, medication, diagnosis, and procedural information) was used to determine a patient’s eligibility for inclusion.

Based on stakeholder feedback and a review of literature and existing quality measures, the following numerator, denominator, and exclusion statements were used in alpha and beta testing:

Denominator: All patients, aged 18 years or older, who received an elective primary THA and/or TKA procedure and do not meet any exclusion criteria.

Figure 1: Denominator Exclusion Criteria

<table>
<thead>
<tr>
<th>Exclusions: Patients who meet any of the criteria below will be excluded from the measurement population</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Who were discharged against medical advice (AMA)</td>
</tr>
<tr>
<td>• Who had more than two THA/TKA procedure codes during the index hospitalization</td>
</tr>
<tr>
<td>• With diagnosis codes for renal insufficiency within the 365 days prior to the THA/TKA procedure</td>
</tr>
<tr>
<td>• With diagnosis codes for chronic atrial fibrillation within the 365 days prior to the THA/TKA procedure</td>
</tr>
<tr>
<td>• With diagnosis codes for cancer within the 365 days prior to the THA/TKA procedure</td>
</tr>
<tr>
<td>• Who received prescription orders for anticoagulant medications 10-90 Days Prior to Surgery and meets the following criteria:</td>
</tr>
<tr>
<td>o Patient who received an Anticoagulant Injection/Infusion</td>
</tr>
<tr>
<td>o Patient who received a tablet (Oral) Anticoagulant order, Quantity &gt; 1</td>
</tr>
<tr>
<td>• With VTE diagnosis code present on admission for index admission</td>
</tr>
<tr>
<td>• With major bleeding diagnosis code present on admission for index admission</td>
</tr>
<tr>
<td>• With diagnosis code for coagulation disorder within the 365 days prior to the THA/TKA procedure</td>
</tr>
<tr>
<td>• Who had additional surgery within 35 days from the elective primary THA/TKA</td>
</tr>
</tbody>
</table>

Numerator: Patients who develop a major bleeding and/or VTE event occurring from the date of the THA/TKA procedure to 35 days postdate of procedure.

• Diagnosed with a VTE event (Deep Vein Thrombosis or Pulmonary Embolism)
• Presence of ICD 10 diagnosis codes for major bleeding during index admission, inpatient hospital encounters, and/or outpatient physician office visits
• Presence of ICD 10 PCS procedure codes for treatment of hemorrhage or hematoma during the index admission, inpatient hospital encounters, and/or outpatient physician office visits
• If the patient received transfusions of ≥1 unit of whole blood or packed cells

Statistical Testing and Analysis:

Alpha Testing
The development team first performed testing to determine the feasibility of implementing this eCQM into existing EHR systems, focusing on the availability of data elements needed for measurement and risk adjustment. Validity testing was conducted from Site 1 to evaluate the agreement between the eCQM and a manual chart review. Chart reviews were performed on a random sample of 236 patients from Site 1, and 30 patients from Site 2. Separate chart reviews were performed on random samples of both patients who met the inclusion criteria, and random samples of patients who met exclusion criteria, to assess the validity and data agreement between manual review and the eCQM. If disagreements between the manual reviewer and the eCQM were found, the particular disagreement would be documented and explored further by the eCQM development team. In some instances, additional EHR review was required to understand and resolve the disagreement. Multiple rounds of review were conducted as measure specifications were fine tuned to resolve disagreements and/or errors.

Beta Testing
Data were extracted from both sites to analyze the type of surgery performed (THA and/or TKA), patient characteristics, and diagnostic and procedural information. Individual clinician group major bleeding rates were risk adjusted by adapting Yale CORE’s Hospital-Level Risk-Standardized Complication Rate risk adjustment model17.
The model utilized by the eCQM development team considers the number of procedures, sex, race, area-level household income (using zip code information), BMI, smoking status, language, and comorbid conditions (Table 1). The team utilized a test-retest approach to test the reliability of the predicted/expected ratios at the clinician group level. Within each clinician group, we randomly selected 50% of the THA/TKA admissions for a test sample, which was used to create the model. We used the remaining random sample of 50% of THA/TKA admissions as the validation sample. We ranked the predicted/expected ratios in the test and validation samples and then estimated the Spearman rank correlation coefficient to correlate the ranking in the test and validation samples. Additionally, the C-Statistic/Area under ROC Curve and the Hosmer-Lemeshow Goodness of Fit test was performed to validate the adequacy of the model.

**Measure Scoring**

The risk adjusted bleeding rate is the ratio of the number of “predicted” events to the number of “expected” events for each clinician group, multiplied by the overall rates calculated in our measure. The measure estimates clinician group-level bleeding rates using hierarchical logistic regression models. This approach models the log-odds of a patient experiencing a bleed using the risk adjustment variables and clinician group-specific intercepts. At the clinician group level, it models the group-specific intercepts as arising from a normal distribution. In this model, the clinician group intercept represents the underlying risk of a bleed for procedures performed by said group, after accounting for the patient risk factors in the risk adjustment model. If there were no differences among clinician groups, then after adjusting for patient risk, the group level intercepts should be identical across all groups.

The risk adjusted VTE and Bleeding rates are derived from a hierarchical binomial logistic regression model, where, for each patient, the outcome is the sum of the VTE outcome (1=yes, 0=no) plus the bleed outcome (1=yes, 0=no); the sum equals 0 (no events), 1 (either VTE or bleed, but not both), or 2 (both VTE and bleed). From this hierarchical binomial logistic regression model, the risk adjusted VTE and Bleeding rates are calculated by taking the (predicted/expected ratios) * (overall sample rate) * 100. Risk-adjustment of the VTE data was not possible due to the low number of events within the total sample. However, by combining the number of VTE and major bleeding events for each clinician group a risk adjusted composite score can be calculated. In our testing, the composite score (combination of VTE and bleeding) for each clinician group is calculated by (predicted/expected ratios) x (overall score) x 100, where the overall score is equal to the (number of VTE and bleeding events) / (Denominator * 2). Clinician groups can be ranked based on these performance metrics by using the predicted/expected ratios or the risk-adjusted rates and scores.

**Results**

The total sample included 17,333 patients from Site 1 (2016-2019) and 11,885 patients from Site 2 (2017-2019). From the total sample, 74.5% of Site 1 patients (n=12,913) and 89.6% of Site 2 patients (n=10,646) met the inclusionary criteria and were included in analysis. The eCQM team evaluated the population of patients included in our test and validation samples and found there were no differences between clinician groups (p = 0.999) or between Sites (p = 0.938). Descriptive statistics of the sample population can be found in Table 1.

**Statistical Testing and Analysis:**

**Alpha Testing**

Overall, the eCQM reliably captured 98% of the data elements required for rate calculation and risk adjustment. Smoking status was the only data element that was captured less than 90% of the time. Refer to Table 2 for more detailed information regarding data element reliability. Validity testing within Site 1 indicated that the manual chart review and the eCQM had excellent agreement by the end of the measure development process. In the final round of testing, the eCQM and chart reviewers agreed on 96% of the cases randomly selected from Site 1 and 93% of cases randomly selected from Site 2. Results of the iterative chart review process can be seen in Table 3. Overall, the results of alpha testing indicated that this eCQM can be readily implemented into Epic, Cerner, and other EHR systems that adopt recommended terminology standards.
Table 1: Descriptive Statistics of Total Sample

<table>
<thead>
<tr>
<th>Variable</th>
<th>Site 1</th>
<th>Site 2</th>
<th>Comorbid Conditions (Cont.)</th>
<th>Site 1</th>
<th>Site 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Admissions</td>
<td>17333</td>
<td>11885</td>
<td>Chronic obstructive pulmonary disease (COPD)</td>
<td>4.68%</td>
<td>3.89%</td>
</tr>
<tr>
<td>Eligible Patients</td>
<td>12914</td>
<td>10643</td>
<td>Decubitus ulcer or chronic skin ulcer</td>
<td>0.16%</td>
<td>0.24%</td>
</tr>
<tr>
<td>Number of Clinician Groups</td>
<td>6</td>
<td>11</td>
<td>Dementia or other specified brain disorders</td>
<td>0.89%</td>
<td>1.99%</td>
</tr>
<tr>
<td>Number of VTE Events</td>
<td>71</td>
<td>22</td>
<td>Diabetes mellitus (DM) or DM complications</td>
<td>11.93%</td>
<td>17.53%</td>
</tr>
<tr>
<td>Number of Bleeding Events</td>
<td>375</td>
<td>519</td>
<td>Dialysis status</td>
<td>0.14%</td>
<td>0.35%</td>
</tr>
<tr>
<td>Hemiplegia paraplegia paralysis</td>
<td></td>
<td></td>
<td></td>
<td>0.73%</td>
<td>0.77%</td>
</tr>
</tbody>
</table>

**Demographics**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Site 1</th>
<th>Site 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean Age</td>
<td>66.05</td>
<td>65.63</td>
</tr>
<tr>
<td>Age ≥ 65 years</td>
<td>59.42%</td>
<td>56.94%</td>
</tr>
<tr>
<td>18 ≤ Age ≤ 65 years</td>
<td>40.58%</td>
<td>43.06%</td>
</tr>
<tr>
<td>Male (%)</td>
<td>42.10%</td>
<td>42.12%</td>
</tr>
<tr>
<td>Median BMI</td>
<td>29.61</td>
<td>30.47</td>
</tr>
<tr>
<td>White</td>
<td>89.89%</td>
<td>68.13%</td>
</tr>
<tr>
<td>Black/African American</td>
<td>3.70%</td>
<td>N/A</td>
</tr>
<tr>
<td>Hispanic</td>
<td>0.51%</td>
<td>N/A</td>
</tr>
<tr>
<td>English as first language</td>
<td>95.55%</td>
<td>92.46%</td>
</tr>
<tr>
<td>Smoker</td>
<td>5.54%</td>
<td>6.25%</td>
</tr>
<tr>
<td>Insurance - % Public</td>
<td>46.79%</td>
<td>56.64%</td>
</tr>
<tr>
<td>Median Income</td>
<td>$72,935</td>
<td>$63,795</td>
</tr>
<tr>
<td>THA/TKA Procedure</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Respiratory/heart/digestive/urinary/other neoplasms</td>
<td>6.18%</td>
<td>5.12%</td>
</tr>
<tr>
<td>THA (%)</td>
<td>45.10%</td>
<td>48.85%</td>
</tr>
<tr>
<td>TKA (%)</td>
<td>54.58%</td>
<td>51.14%</td>
</tr>
<tr>
<td>Comorbid Conditions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trauma</td>
<td>2.03%</td>
<td>2.40%</td>
</tr>
<tr>
<td>Vascular or circulatory disease</td>
<td>3.77%</td>
<td>3.53%</td>
</tr>
<tr>
<td>Vertebral fractures without spinal cord injury</td>
<td>0.93%</td>
<td>1.45%</td>
</tr>
<tr>
<td>Major complications of medical care and trauma</td>
<td>10.34%</td>
<td>8.79%</td>
</tr>
</tbody>
</table>

*Rates of osteoarthritis in Site 2 are validated. The difference in osteoarthritis rates is attributable to Site 1 rates being recorded upon index admission prior to the procedure, while the Site 2 rates are recorded upon discharge, following the procedure.

Table 2: Availability of Data Elements for the Risk Adjustment Model

<table>
<thead>
<tr>
<th>Variable</th>
<th>Site 1 Frequency</th>
<th>Site 1 % Missing</th>
<th>Site 2 Frequency</th>
<th>Site 2 % Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>INSURANCE TYPE</td>
<td>17316</td>
<td>0.10%</td>
<td>11884</td>
<td>0.01%</td>
</tr>
<tr>
<td>BODY MASS INDEX (BMI)</td>
<td>17211</td>
<td>0.70%</td>
<td>11876</td>
<td>0.08%</td>
</tr>
<tr>
<td>PRIMARY LANGUAGE</td>
<td>17175</td>
<td>0.91%</td>
<td>11867</td>
<td>0.15%</td>
</tr>
<tr>
<td>SMOKING STATUS</td>
<td>17079</td>
<td>1.47%</td>
<td>10347</td>
<td>12.94%</td>
</tr>
<tr>
<td>ZIPCODE</td>
<td>17308</td>
<td>0.14%</td>
<td>11869</td>
<td>0.13%</td>
</tr>
<tr>
<td>SEX</td>
<td>17333</td>
<td>0.00%</td>
<td>11878</td>
<td>0.06%</td>
</tr>
<tr>
<td>RACE</td>
<td>16946</td>
<td>0.23%</td>
<td>11880</td>
<td>0.04%</td>
</tr>
<tr>
<td>ADMIT AGE</td>
<td>17333</td>
<td>0.00%</td>
<td>11885</td>
<td>0.00%</td>
</tr>
<tr>
<td>CONDITION</td>
<td>None of the cases had all NULL values for any of the condition columns</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Validity Testing Results

<table>
<thead>
<tr>
<th>Chart Review Round</th>
<th>N</th>
<th>Kappa Value</th>
<th>Kappa Lower 95% CL</th>
<th>Kappa Upper 95% CL</th>
<th>% AGREEMENT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Site 1 (1)</td>
<td>72</td>
<td>0.758</td>
<td>0.641</td>
<td>0.875</td>
<td>81.90%</td>
</tr>
<tr>
<td>Site 1 (2)</td>
<td>64</td>
<td>0.667</td>
<td>0.528</td>
<td>0.805</td>
<td>75.00%</td>
</tr>
<tr>
<td>Site 1 (3)</td>
<td>100</td>
<td>0.929</td>
<td>0.861</td>
<td>0.998</td>
<td>96.00%</td>
</tr>
<tr>
<td>Site 2 (1)</td>
<td>30</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>93.33%</td>
</tr>
</tbody>
</table>
**Beta Testing**

Site 1 (n=12,914) had an unadjusted major bleeding rate of 2.90% with clinician group performance ranging between (2.54%– 4.76%) and a VTE rate of 0.55% (0.20%– 0.84%). Site 2 (n=10,643) had an unadjusted bleeding rate of 4.86% (2.17%– 39.66%) and a VTE rate of 0.21% (0.00%– 2.65%). The unadjusted data shown in Table 4 also indicates that there was substantial variation in unadjusted bleeding and VTE rates between different clinician groups regardless of site.

However, the measurement of interest for each clinician group was the ratio of the number of “predicted” to the number of “expected” events. Both the predicted and expected numerator events were adjusted for age, gender, type of surgery (THA/TKA), insurance, race, household income, English as primary language, smoking status, body mass index and comorbidities. Table 5 below shows the predicted/expected ratios and 95% confidence intervals for the 17 clinician groups (blinded). Only the Major Bleeding and Composite predicted/expected ratios are shown since the development team was unable to risk-adjust the VTE data due to the low number of events within the total sample. The calculated rates/scores for all 17 clinician groups are shown in Table 6. The overall adjusted bleeding rate was 3.39%/3.398% and ranged between 1.987% – 5.656% for the Test and Validation Samples. The overall VTE/Bleeding composite score was 1.892/1.923 and ranged between 1.148 – 3.184 for the Test and Validation Samples.

### Table 4: Unadjusted Major Bleeding and VTE rates for the 17 clinician groups in the test and validation samples

<table>
<thead>
<tr>
<th>Site</th>
<th>Clinician Group</th>
<th>Surgeries (N)</th>
<th>Bleeding Events (N)</th>
<th>Unadjusted Bleeding Rate</th>
<th>VTE Events (N)</th>
<th>Unadjusted VTE Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td>4920</td>
<td>132</td>
<td>2.683%</td>
<td>29</td>
<td>0.589%</td>
</tr>
<tr>
<td>1</td>
<td>B</td>
<td>593</td>
<td>27</td>
<td>4.553%</td>
<td>5</td>
<td>0.843%</td>
</tr>
<tr>
<td>1</td>
<td>C</td>
<td>3151</td>
<td>80</td>
<td>2.539%</td>
<td>24</td>
<td>0.762%</td>
</tr>
<tr>
<td>1</td>
<td>D</td>
<td>180</td>
<td>0</td>
<td>0.000%</td>
<td>1</td>
<td>0.556%</td>
</tr>
<tr>
<td>1</td>
<td>E</td>
<td>1029</td>
<td>49</td>
<td>4.762%</td>
<td>6</td>
<td>0.583%</td>
</tr>
<tr>
<td>1</td>
<td>F</td>
<td>3041</td>
<td>87</td>
<td>2.861%</td>
<td>6</td>
<td>0.197%</td>
</tr>
<tr>
<td>2</td>
<td>G</td>
<td>148</td>
<td>8</td>
<td>5.405%</td>
<td>0</td>
<td>0.000%</td>
</tr>
<tr>
<td>2</td>
<td>H</td>
<td>319</td>
<td>14</td>
<td>4.389%</td>
<td>1</td>
<td>0.313%</td>
</tr>
<tr>
<td>2</td>
<td>I</td>
<td>58</td>
<td>23</td>
<td>39.655%</td>
<td>0</td>
<td>0.000%</td>
</tr>
<tr>
<td>2</td>
<td>J</td>
<td>2438</td>
<td>53</td>
<td>2.174%</td>
<td>6</td>
<td>0.246%</td>
</tr>
<tr>
<td>2</td>
<td>K</td>
<td>459</td>
<td>31</td>
<td>6.754%</td>
<td>1</td>
<td>0.218%</td>
</tr>
<tr>
<td>2</td>
<td>L</td>
<td>3456</td>
<td>155</td>
<td>4.485%</td>
<td>4</td>
<td>0.116%</td>
</tr>
<tr>
<td>2</td>
<td>M</td>
<td>151</td>
<td>30</td>
<td>19.868%</td>
<td>4</td>
<td>2.649%</td>
</tr>
<tr>
<td>2</td>
<td>N</td>
<td>830</td>
<td>65</td>
<td>7.831%</td>
<td>1</td>
<td>0.120%</td>
</tr>
<tr>
<td>2</td>
<td>O</td>
<td>406</td>
<td>18</td>
<td>4.433%</td>
<td>1</td>
<td>0.246%</td>
</tr>
<tr>
<td>2</td>
<td>P</td>
<td>474</td>
<td>55</td>
<td>11.603%</td>
<td>1</td>
<td>0.211%</td>
</tr>
<tr>
<td>2</td>
<td>Q</td>
<td>1923</td>
<td>66</td>
<td>3.432%</td>
<td>3</td>
<td>0.156%</td>
</tr>
</tbody>
</table>

### Tables 5: Predicted/Expected ratios for the 17 clinician groups (blinded) for the test and validation samples

<table>
<thead>
<tr>
<th>Site</th>
<th>Clinician Group</th>
<th>N</th>
<th>Risk-Adjusted Major Bleeding</th>
<th>Risk-Adjusted Composite Score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Test</td>
<td>95% CI</td>
</tr>
<tr>
<td>1</td>
<td>A</td>
<td>4920</td>
<td>0.839</td>
<td>0.580-1.216</td>
</tr>
<tr>
<td>1</td>
<td>B</td>
<td>593</td>
<td>1.485</td>
<td>0.965-2.287</td>
</tr>
<tr>
<td>1</td>
<td>C</td>
<td>3151</td>
<td>0.779</td>
<td>0.514-1.183</td>
</tr>
<tr>
<td>1</td>
<td>D</td>
<td>180</td>
<td>0.674</td>
<td>0.296-1.534</td>
</tr>
<tr>
<td>1</td>
<td>E</td>
<td>129</td>
<td>1.115</td>
<td>0.757-1.643</td>
</tr>
<tr>
<td>1</td>
<td>F</td>
<td>3041</td>
<td>0.909</td>
<td>0.615-1.347</td>
</tr>
<tr>
<td>2</td>
<td>G</td>
<td>148</td>
<td>1.059</td>
<td>0.543-2.069</td>
</tr>
<tr>
<td>2</td>
<td>H</td>
<td>319</td>
<td>0.856</td>
<td>0.493-1.487</td>
</tr>
<tr>
<td>2</td>
<td>I</td>
<td>58</td>
<td>1.377</td>
<td>0.994-1.911</td>
</tr>
<tr>
<td>2</td>
<td>J</td>
<td>2438</td>
<td>0.523</td>
<td>0.301-0.911</td>
</tr>
<tr>
<td>2</td>
<td>K</td>
<td>459</td>
<td>0.929</td>
<td>0.589-1.467</td>
</tr>
<tr>
<td>2</td>
<td>L</td>
<td>3456</td>
<td>0.797</td>
<td>0.566-1.223</td>
</tr>
<tr>
<td>2</td>
<td>M</td>
<td>151</td>
<td>1.189</td>
<td>0.818-1.728</td>
</tr>
<tr>
<td>2</td>
<td>N</td>
<td>830</td>
<td>1.490</td>
<td>1.124-1.977</td>
</tr>
<tr>
<td>2</td>
<td>O</td>
<td>406</td>
<td>0.996</td>
<td>0.578-1.717</td>
</tr>
<tr>
<td>2</td>
<td>P</td>
<td>474</td>
<td>1.350</td>
<td>0.985-1.852</td>
</tr>
<tr>
<td>2</td>
<td>Q</td>
<td>1923</td>
<td>0.940</td>
<td>0.642-1.378</td>
</tr>
</tbody>
</table>

740
Overall, when evaluating the risk-adjustment model, our results demonstrate strong reliability across the test and validation samples. As noted in Tables 5 and 6, the test and validation samples give similar rankings of the 17 clinician groups with respect to the predicted/expected ratios with a Spearman rank Correlation of .900 and .946 for Bleeding and the Composite Score respectively.

For major bleeding alone we estimated the intraclass correlation (ICC) between clinician groups and the ICC value was .054 meaning there is some variability across clinician groups as noted by the predicted/expected ratios in Table 5. When taking the VTE component into account within the composite score the ICC value is .041, indicating some variability between clinician groups.

**Adequacy of Model**

The resulting C-statistics for the test and validation sample were the following for Major Bleeding and the Composite Score: 0.796/0.779 and 0.782/0.773. Additionally, Hosmer-Lemeshow tests were conducted on the test and validation samples (p-values test/validation: major bleeding model 0.978/0.214 and composite score model 0.345/0.607).

**Discussion:**

This study reports on the initial testing of an eCQM for VTE/major bleeding across clinical groups within two healthcare systems for people undergoing THA and/or TKA. With regards to the feasibility of implementing this eCQM, the research team found that the required data elements for rate calculation and risk adjustment were reliably captured across both sites and clinician groups, missing less than 2% for most data elements. Since the vast majority of data elements were captured, the eCQM team did not believe that the missing data would bias results. Performance wise, beta testing indicated an overall unadjusted bleeding rate of 3.8% and a VTE rate of 0.4%, which is in line with the rates that are currently in the literature, ranging from 0.2% to 6.8% and 0.2% to 3.7% respectively. Upon initial observation, relatively large variations in both major bleeding and VTE rates were noted between clinician groups. These differences may indicate that inconsistent approaches to periprocedural anticoagulant prescribing practices exist between different clinician groups and health systems. This is unsurprising as the guidelines outlined by the American Academy of Orthopaedic Surgeons and the American College of Chest Physicians indicate that while VTE prophylaxis is recommended, no specific recommended regimen nor pharmacological agent exists. Even though anticoagulant prescribing practices may vary, variability within the patient population of each clinician group may additionally be contributing to the differences observed between clinician groups. However, the risk-adjusted results, which accounts for differences in patient population, for both the major bleeding and component scores indicate that clinically significant variations in performance can be seen between the 17 clinician groups. Statistical testing evaluating the model itself was conducted utilizing C-statistics and Hosmer-Lemeshow tests. The results indicated that our predictive model showed good discrimination of who would experience a VTE and/or major bleeding event. Additionally, there were no statistically significant differences between the predicted and expected results for either the major bleeding or composite models.

### Table 6: Risk Adjusted Rates and Scores for the 17 clinician groups

<table>
<thead>
<tr>
<th>Site</th>
<th>Clinician Group</th>
<th>N</th>
<th>Risk-Adjusted Bleeding Rate</th>
<th>Risk-Adjusted Composite Score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Test</td>
<td>Validation</td>
</tr>
<tr>
<td>1</td>
<td>A</td>
<td>4920</td>
<td>3.187%</td>
<td>3.357%</td>
</tr>
<tr>
<td>1</td>
<td>B</td>
<td>593</td>
<td>5.637%</td>
<td>5.365%</td>
</tr>
<tr>
<td>1</td>
<td>C</td>
<td>3151</td>
<td>2.959%</td>
<td>3.203%</td>
</tr>
<tr>
<td>1</td>
<td>D</td>
<td>180</td>
<td>2.558%</td>
<td>2.610%</td>
</tr>
<tr>
<td>1</td>
<td>E</td>
<td>1029</td>
<td>4.231%</td>
<td>4.662%</td>
</tr>
<tr>
<td>1</td>
<td>F</td>
<td>3041</td>
<td>3.453%</td>
<td>3.440%</td>
</tr>
<tr>
<td>2</td>
<td>A</td>
<td>148</td>
<td>4.020%</td>
<td>3.941%</td>
</tr>
<tr>
<td>2</td>
<td>B</td>
<td>319</td>
<td>3.248%</td>
<td>3.396%</td>
</tr>
<tr>
<td>2</td>
<td>C</td>
<td>58</td>
<td>5.229%</td>
<td>4.418%</td>
</tr>
<tr>
<td>2</td>
<td>D</td>
<td>2438</td>
<td>1.987%</td>
<td>2.068%</td>
</tr>
<tr>
<td>2</td>
<td>E</td>
<td>459</td>
<td>3.527%</td>
<td>4.480%</td>
</tr>
<tr>
<td>2</td>
<td>F</td>
<td>3456</td>
<td>3.024%</td>
<td>3.053%</td>
</tr>
<tr>
<td>2</td>
<td>G</td>
<td>151</td>
<td>4.512%</td>
<td>4.422%</td>
</tr>
<tr>
<td>2</td>
<td>H</td>
<td>830</td>
<td>5.656%</td>
<td>4.900%</td>
</tr>
<tr>
<td>2</td>
<td>I</td>
<td>406</td>
<td>3.781%</td>
<td>3.593%</td>
</tr>
<tr>
<td>2</td>
<td>J</td>
<td>474</td>
<td>5.125%</td>
<td>5.062%</td>
</tr>
<tr>
<td>2</td>
<td>K</td>
<td>1923</td>
<td>3.567%</td>
<td>3.338%</td>
</tr>
</tbody>
</table>

Adjusted Bleeding Rate = (predicted/expected ratios) x (overall rate) where overall rate = 0.038
Composite Score = (predicted/expected ratios) x (overall rate) x 100 where overall rate = 0.021
To our knowledge, this is the first measure that evaluates both the bleeding and VTE rate following TJA. The development team believes that this eCQM could enable a better understanding of how a specific clinician group’s post-operative anticoagulation and bleeding outcomes compare with their peers. Tying in monetary incentives through the MIPS participation pathway of the Quality Payment Program (QPP), orthopedic clinician groups will be motivated to improve their practices, ultimately driving quality improvement within the specialty of orthopedic surgery. Importantly, as an eCQM, measure scores can be calculated at any time, giving clinicians the ability to evaluate their performance at regular intervals as opposed to claims-based measures that are only published every two years.

Several limitations were acknowledged during eCQM testing. One limitation is that this eCQM can only be used within integrated health care networks that utilize an enterprise EHR system. For example, if a patient receives care from multiple health sites with different or unconnected EHR systems, a patient’s medical records could be scattered across several independent databases. Similarly, another limitation is the lack of consistent policies surrounding EHR documentation. At the Cerner Test Site, we discovered that some data elements were documented consistently (mandated) while others were optional due differences in policy (smoking status). With our risk-adjustment modeling heavily relying on data element availability, the ability to risk-adjust within other healthcare systems depends on current and future policies enacted by healthcare systems. Throughout our testing process, the eCQM team discovered that documentation procedures vary widely even between clinician groups within the same site as each clinician group may have different thresholds/protocols for the documentation or treatment of certain events. Another limitation of the measure is that bleeding and VTE harm are weighted equally because the current version of the CMS Measure Authoring Tool (MAT) is unable to assign unequal weights. VTE events are much less common than major bleeding events, but they pose a significantly greater risk of serious harm/death to the patient. In the future, once enhancements are made to MAT, the BWH team will ideally use the Harm Weights established by the AHRQ PSI 90 composite measure where Perioperative Hemorrhage or Hematoma Rate is assigned a harm weight of 0.0570 and PSI 12 Perioperative Pulmonary Embolism or Deep Vein Thrombosis Rate is assigned a harm rate of 0.1557. This would result in a Harm Weight Ratio equivalent to 2.7 for VTE events, and would be the score given to patients who experience a VTE event compared to the score of 1 for a Major Bleeding event. Lastly, the eCQM has only been tested within clinician groups at two large academic healthcare care systems with two different EHR vendor systems. Testing at additional sites is needed to fully understand the feasibility and meaningfulness of this eCQM. However, we conclude that this appears to be a useful measure for these procedures especially as it includes the competing risks of clotting and bleeding.

The BWH eCQM team has submitted this measure to the 2021 CMS Measures under consideration (MUC) List and plans to obtain NQF endorsement to facilitate implementation into the CMS’ QPP. If implemented, this VTE and Major Bleeding eCQM will enhance the MIPS pathway by providing clinicians with a minimally burdensome method to monitor the impact of their anticoagulation practices while minimizing provider time spent collecting and submitting data to CMS.

References
2. CMS, Comprehensive Care for Joint Replacement Model, CMS.gov.
15. American Academy of Orthopaedic Surgeons Clinical Practice Guideline on Preventing Thromboembolic Disease in Patients Undergoing Elective Hip and Knee Arthroplasty
Identifying Sleep-Related Factors Associated with Cognitive Function in a Hispanics/Latinos Cohort: A Dual Random Forest Approach

Xiaojin Li, Ph.D.1, Licong Cui, Ph.D.1, Fei Wang, Ph.D.2, Paul E. Schulz, M.D.1, Guo-Qiang Zhang, Ph.D.1
1The University of Texas Health Science Center at Houston, Houston, TX 77030
2Department of Population Health Sciences, Weill Cornell Medicine, New York, NY 10065

Abstract
Disordered sleep is associated with poor cognitive function and cognitive decline. However, little is known regarding the association of sleep-related factors with cognitive function in underrepresented cohorts such as the Hispanic/Latino population. Leveraging the National Sleep Research Resource, one of the most comprehensive collections of sleep studies, we identified a Hispanic/Latino cohort of 1,031 lower cognitive function cases and 2,062 normal controls. We developed a novel dual random forest (DRF) approach to discriminate cases against controls for estimating the potential impact of sleep-related variables related to the decline of cognitive function. Several important sleep-related factors were identified which may be associated with cognitive function in the Hispanics/Latinos cohort, such as heart rate, sleep duration, trouble falling asleep, and apnea/hypopnea index, which are consistent with existing research findings. Our DRF approach is effective in validating the association between disordered sleep and cognitive decline in this unique minority population.

1 Introduction
Sleep is a growing and under-appreciated determinant of health and well-being1–4. Disordered sleep-wake cycles and sleep duration (quantity), efficiency (quality), apnea, latency, and stages (rapid eye movement [REM] and non-rapid eye movement [NREM]) have been considered as important markers of sleep architecture associated with cardiovascular disease2,3,5,6, depression7,8, diabetes9,10, hypertension11, and lately, Alzheimer’s disease and related dementias (ADRD)12–14. Alzheimer’s disease (AD), a debilitating neurological condition that impairs memory, thought processes, and function, has become a public health crisis15. AD affects about 5.7 million Americans16,17 and one in ten people (10%) aged 65 and older. As AD occurs primarily among older adults, the number of individuals with ADRD is projected to double by 2050 due to the aging American population. The cost of AD to society is substantial as people with ADRD require significant expenditures for health care, intensive long-term services, and support. The total costs of health care for individuals with AD in 2017 is estimated at almost $260 billion18.

The Mini Mental State Examination (MMSE) is a gold-standard measurement in clinical and research settings for estimating global cognitive function status and screening for dementia19. It consists of 30-point questions assessing orientation, attention, memory, language, and visuospatial skills19, such as repeating lists of words, the time and place of the test, arithmetic, language use and comprehension, and basic motor skills20. The Six-Item Screener (SIS) is a simple test derived from the MMSE (consisting of 3-item recall and 3-item temporal orientation) for identifying subjects with cognitive impairment to meet the needs of being brief, sensitive, and easily remembered in time-limited situations such as the emergency department21,22 and primary care setting21.

A large number of prior studies has focused on the relationship between disordered sleep duration and poor cognitive function or cognitive decline using MMSE or similar measures4,13,24–30. For example, Terri et al. reported modest cross-sectional associations of wake after sleep onset and self-reported long sleep with cognition decline among the older community25,26. Alberto et al. indicated long sleep duration was associated with worse MMSE performance in an elderly community cohort from the Northern Manhattan Study27. Mari et al. reported that short-term memory appears to decline with disturbed sleep (day and overnight shifts)28. Obstructive sleep apnea (OSA) was found to be associated with a greater risk of mild cognitive impairment or dementia13,29,30.

In this paper, we leverage a Hispanics/Latinos cohort in the National Sleep Research Resource (NSRR) to investigate the potential sleep-related factors that can differentiate lower cognitive function (LCF) group from the normal control (NC) group according to the SIS, using a novel dual random forest (DRF) approach. The main contributions of this work are: (1) we presented a methodology that is the first time to derive discriminative sleep-related factors for
cognitive function and classify different levels of cognitive function in a Hispanics/Latinos cohort; (2) we leveraged a rich collection of de-identified clinical data from NSRR; and (3) we obtained representative sleep-related factors consistent with the results of existing studies, demonstrating the validity of our approach.

2 Background

2.1 National Sleep Research Resource (NSRR)

The NSRR (R24HL114473; 2013-19)\textsuperscript{31–33} offers free and open web access to large collections of de-identified, well-annotated national repository of sleep data, including polysomnograms (PSGs) which are linked to risk factor and outcome data for participants in major NIH observational studies and clinical trials. This national repository of sleep data, the first of its kind, is significant because neurophysiological data has not been previously made available at such a large and systematic scale. The NSRR provides opportunities for investigators to address critical questions about the impact of sleep disorders on important clinical outcomes, thereby enhancing clinical and translational work in human sleep medicine and physiology. NSRR further provides users with a suite of tools to facilitate data exploration and data visualization. Launched in 2014, more than 10 billion data files, totaling over 500 terabytes, were downloaded by 4,000 registered researchers around the world from the NSRR portal (sleepdata.org). Recently, NSRR has been designated one of the Open Domain-Specific Data Sharing Repositories by the Trans-NIH BioMedical Informatics Coordinating Committee.

2.2 The Hispanic Community Health Study / Study of Latinos (HCHS/SOL)

The Hispanic Community Health Study / Study of Latinos (HCHS/SOL)\textsuperscript{34}, one of the datasets integrated in the NSRR, is a multi-center epidemiological study of Hispanic/Latino populations consisting of more than 16,000 self-identified Hispanics/Latinos. The aims of HCHS/SOL are to identify the role of acculturation in the prevalence and development of specific diseases, and to identify risk factors playing a protective or harmful role in Hispanics/Latinos. HCHS/SOL was sponsored by the National Heart, Lung, and Blood Institute (NHLBI) and six other institutes, centers, and offices of the National Institutes of Health (NIH) contributed to the first phase of the project as well.

3 Methods

We develop a computational framework to investigate sleep-related factors that could be discriminative regarding cognitive function (as shown in Figure 1). This framework consists of five steps: (1) extracting variables from dataset and grouping subjects based on SIS, (2) identifying covariates using the first random forest model (as RF-A in Figure 1), (3) matching and selecting subjects using similarity score matching (SSM), (4) applying bivariate analysis on sleep-related clinical variables, and (5) discriminating LCF cases and NCs using the second random forest model (as RF-B in Figure 1). Steps 1-3 create a cohort that contains subjects from both the LCF and NC groups such that both groups have similar covariate distributions. Step 4 removes variables that are not statistically relevant to cognitive function and reduces data dimensionality. Step 5 generates the feature importance of selected candidates.

![Figure 1: The overall framework to investigate discriminative sleep-related factors associated with cognitive function.](image)

3.1 Data description

We leverage the HCHS/SOL dataset in NSRR, which contains large collections of de-identified clinical data elements from over 16,000 Hispanics/Latinos. We extract 230 clinical data elements/variables from 15 categories including social demographic (e.g., age, gender, and education status), blood pressure (e.g., overall ankle branchial index), dental (e.g., periodontitis severity), global physical activity (e.g., total physical activity per week), anthropometry
(e.g., body mass index and waist to hip ratio), laboratory measures (e.g., insulin fasting), quality of life (e.g., 10-Item state trait anxiety inventory), clinical characteristics (e.g., alcohol use), medications (e.g., highest reported frequency of sleeping medication usage), respiratory (e.g., bronchodilator responsive by 15% change), health survey (e.g., aggregate physical health score), medical history (e.g., stroke and diabetes), sleep monitoring (e.g., mean peripheral capillary oxygen saturation [SpO2] desaturation), sleep questionnaire (e.g., feel sleepy during the day), and sleep summary (e.g., average sleep duration and insomnia severity index). As in total, 8,216 subjects in the age group of 19 to 76 years are used in further covariate analysis and subject selection. The subjects are categorized into two groups, LCF and NC, based on SIS score, which are dichotomized to 4 or less, or above 4 based on previous studies of neurocognitive impairment\textsuperscript{35}.

3.2 Covariate analysis

To focus on the relationship between cognitive function and sleep-related factors, we need to reduce the confounding effects of other covariates that correlate with the incidence of LCF, such as demographic background and certain risk factors. We randomly select subjects from NC group with the same number of subjects in the LCF group, and apply the random forest model to the selected subjects using extracted clinical variables as features. We run this process 10 times and select the variables with high average feature importance (the higher the value, the more important the feature) as covariates. Also, we manually add confounders mentioned in the previous studies\textsuperscript{36} as covariates, such as hypertension, stroke, and body mass index (BMI).

3.3 Subject matching and selection

Among 8,216 subjects, we select NCs that match with LCF cases with a 2:1 ratio in terms of covariates to cognitive function. Inspired by the cohort matching method used in recent machine learning studies\textsuperscript{37}, we apply a similar idea to perform a logistic regression based SSM to reduce or eliminate selection bias in observational studies by balancing covariates between the case (i.e., LCF case) and control group (i.e., NC). Therefore, the analysis results of variables are independent of the outcome conditioned on the confounders\textsuperscript{38}. There are two steps in SSM to match NC to the LCF case. First, we obtain an estimation for the similarity scores using logistic regression. Then we match each LCF case to one or more NCs based on similarity scores using caliper matching, which finds the first 2 NCs that fall within a predefined distance threshold between similarity scores. The criterion for selecting the threshold is to make the covariate statistics of the case group and the control group as similar as possible. The threshold tuning is to ensure our threshold is small enough such that we can get close matches.

3.4 Bivariate analysis

Bivariate analysis is the simultaneous analysis of two variables\textsuperscript{39}. It studies the empirical relationship between two variables, such as whether there is an association and the strength of this association, or the differences and the significance of these differences between two variables\textsuperscript{39}. To remove variables that are not statistically relevant to cognitive function and to reduce data dimensionality, we perform bivariate analysis between each sleep-related variable and SIS score with selected subjects and generate the sleep factor candidates (with p-value < 0.1). Then we feed those candidates to the random forest model. For numerical variables in our bivariate analysis, we use Kendall rank correlation coefficient\textsuperscript{40}, which is a statistic used to measure the association between two measured quantities. For categorical variables, we perform Kruskal-Wallis H test\textsuperscript{41,42}, which is the one-way analysis of variance (ANOVA).

3.5 Random forest

Random forest (RF) is an ensemble learning method for classification or regression by constructing a group of decision trees. The overall output for a specific sample is obtained by weighted combination of the output generated from each decision tree in RF. The weight of each tree is adjusted according to misclassification and out-of-bag measures. There are four steps to build the random forest with the technique of bootstrap aggregating (bagging)\textsuperscript{43}, which is an ensemble method to reduce the variance without increasing the bias for decision tree algorithms (as shown in Figure 2). Given a training set, the first step is to generate several sub-training sets by selecting a random number of observations and features. Steps 2 and 3 are to build and train each decision tree with generated sub-training set and apply an out-of-bag error estimate, which obtains an unbiased estimate of the test set error internally and reduces the overfitting in a
random forest, and there is no need to use further cross-validation\textsuperscript{43}. These three steps are repeated multiple times to build several decision trees. In the last step, the output for unknown samples can be made by averaging the output from all the individual decision trees.

Compared to other traditional classifiers, such as K-nearest neighbor, support vector machine, artificial neural network, RF has the following four advantages. (1) Adaptability: RF estimates the importance of variables and provides a way for tuning with additional training data by assigning different weights for each decision tree; (2) Scalability: RF can handle thousands of input variables and work efficiently on large datasets; (3) Robustness: RF can balance error in datasets with unbalanced class population\textsuperscript{44}; (4) High interpretability: We could use the trained model to gain insight on our data via the learned relationship between feature and response variables (e.g., feature importance). Thus, RF is one of the most popular machine learning methods in feature selection and binary/multi-class classification problems.

4 Experiments and Results

4.1 Covariates and similarity score matching

We obtained top 15 variables (including 11 non-sleep-related and 4 sleep-related variables) and selected 7 covariates based on the feature importance (>0.3, as shown in the Table 1), which was generated by the first RF model with clinical data elements from 15 categories in the HCHS/SOL dataset. We also manually added 5 covariates, including hypertension, stroke, smoke status, alcohol status, and BMI, based on the findings of existing studies\textsuperscript{36}.

To create our training dataset for the RF model, we selected NCs from HCHS/SOL by using similarity score matching to reduce the effect and bias of covariates. After matching, we obtained 1,031 LCF cases and 2,062 NCs at a 1:2 ratio. The summary of the statistics on covariates before and after the matching is shown in Table 2.

4.2 Bivariate and random forest analysis

We performed bivariate analysis based on 124 sleep-related variables and obtained 47 candidates with p-value < 0.1 for further RF analysis. We measured the discrimination based on the accuracy and area under the receiver operating characteristic (ROC) curve (AUROC), two important evaluation metrics for checking a classification model’s performance\textsuperscript{45}, to classify LCF cases and NCs. Table 3 shows the performance with different settings of the number of trees in RF with the metric of Gini impurity. To compute the feature importance, we used the number of trees as 1,000 since it achieved the highest AUROC.

The feature importance of all 47 candidates are shown in Table 4. We also applied our analytical framework to the data source in terms of different genders. There were 542 LCF cases and matched 1,084 NCs for female-specific analysis. For males, we obtained 489 LFC cases and 978 NCs. Figure 3 shows the top 20 important sleep-related variables for females and males, respectively.

Table 1: The top 15 variables obtained from the first random forest model.

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>FI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Education Status</td>
<td>1.0000</td>
</tr>
<tr>
<td>Age</td>
<td>0.7593</td>
</tr>
<tr>
<td>Diet Score from 1st JAMA paper</td>
<td>0.6754</td>
</tr>
<tr>
<td>Yearly Household Income</td>
<td>0.5933</td>
</tr>
<tr>
<td>Employment Status</td>
<td>0.5543</td>
</tr>
<tr>
<td>Gender</td>
<td>0.3832</td>
</tr>
<tr>
<td>10-Item State Trait Anxiety Inventory</td>
<td>0.3794</td>
</tr>
<tr>
<td>Angioplasty/bypass</td>
<td>0.2846</td>
</tr>
<tr>
<td>Heart Attack</td>
<td>0.2769</td>
</tr>
<tr>
<td>Waist to Hip Ratio</td>
<td>0.2579</td>
</tr>
<tr>
<td>Mean SpO2</td>
<td>0.2230</td>
</tr>
<tr>
<td>Angina</td>
<td>0.1935</td>
</tr>
<tr>
<td>Apnea/Hypopnea Index (all desats)</td>
<td>0.1778</td>
</tr>
<tr>
<td>Total Time spent in Apnea/Hypopnea event (all desats) (hours)</td>
<td>0.1714</td>
</tr>
<tr>
<td>Standard Deviation SpO2</td>
<td>0.1653</td>
</tr>
</tbody>
</table>

Notes: FI: Feature Importance.
Table 2: Statistics on covariates before and after the similarity score matching.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Variable Value</th>
<th>Before Matching</th>
<th>After Matching</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>LCF Group</td>
<td>NC Group</td>
</tr>
<tr>
<td></td>
<td></td>
<td>LCF Group</td>
<td>NC Group</td>
</tr>
<tr>
<td>Education Status (%)</td>
<td>No high school diploma or GED</td>
<td>62.35</td>
<td>39.35</td>
</tr>
<tr>
<td></td>
<td>At most a High school diploma/GED</td>
<td>17.27</td>
<td>22.07</td>
</tr>
<tr>
<td></td>
<td>Greater than high school (or GED) education</td>
<td>20.23</td>
<td>38.51</td>
</tr>
<tr>
<td></td>
<td></td>
<td>60.23</td>
<td>59.92</td>
</tr>
<tr>
<td></td>
<td></td>
<td>18.33</td>
<td>19.79</td>
</tr>
<tr>
<td></td>
<td></td>
<td>21.44</td>
<td>21.29</td>
</tr>
<tr>
<td>Age</td>
<td></td>
<td>58.04 ± 8.15</td>
<td>54.91 ± 3.16</td>
</tr>
<tr>
<td></td>
<td></td>
<td>57.44 ± 7.98</td>
<td>51.19 ± 7.91</td>
</tr>
<tr>
<td>Diet Score</td>
<td>Less than $10,000</td>
<td>23.59</td>
<td>15.08</td>
</tr>
<tr>
<td></td>
<td>$10,001 - $15,000</td>
<td>19.38</td>
<td>16.90</td>
</tr>
<tr>
<td></td>
<td>$15,001 - $20,000</td>
<td>12.66</td>
<td>13.18</td>
</tr>
<tr>
<td></td>
<td>$20,001 - $25,000</td>
<td>8.67</td>
<td>10.70</td>
</tr>
<tr>
<td></td>
<td>$25,001 - $29,999</td>
<td>6.17</td>
<td>7.43</td>
</tr>
<tr>
<td></td>
<td>$30,000 - $40,000</td>
<td>7.58</td>
<td>12.96</td>
</tr>
<tr>
<td></td>
<td>$40,001 - $50,000</td>
<td>3.13</td>
<td>6.39</td>
</tr>
<tr>
<td></td>
<td>$50,001 - $75,000</td>
<td>3.13</td>
<td>5.41</td>
</tr>
<tr>
<td></td>
<td>$75,001 - $100,000</td>
<td>0.94</td>
<td>2.31</td>
</tr>
<tr>
<td></td>
<td>More than $100,000</td>
<td>0.47</td>
<td>1.99</td>
</tr>
<tr>
<td>Yearly Household Income (%)</td>
<td></td>
<td>55.23</td>
<td>62.44</td>
</tr>
<tr>
<td></td>
<td></td>
<td>52.57</td>
<td>50.34</td>
</tr>
<tr>
<td>Employment Status (%)</td>
<td>Not Employed</td>
<td>25.94</td>
<td>14.22</td>
</tr>
<tr>
<td></td>
<td>Retired</td>
<td>41.56</td>
<td>36.03</td>
</tr>
<tr>
<td></td>
<td>Part-time</td>
<td>9.61</td>
<td>15.50</td>
</tr>
<tr>
<td></td>
<td>Full-time</td>
<td>21.25</td>
<td>33.17</td>
</tr>
<tr>
<td>Gender (%)</td>
<td>Female</td>
<td>59.23</td>
<td>62.44</td>
</tr>
<tr>
<td></td>
<td>Male</td>
<td>44.77</td>
<td>37.56</td>
</tr>
<tr>
<td>Anxiety Inventory</td>
<td></td>
<td>18.63 ± 6.53</td>
<td>17.10 ± 5.92</td>
</tr>
<tr>
<td></td>
<td></td>
<td>18.43 ± 6.37</td>
<td>18.25 ± 6.43</td>
</tr>
<tr>
<td>Hypertension (%)</td>
<td>Yes</td>
<td>47.81</td>
<td>41.67</td>
</tr>
<tr>
<td></td>
<td></td>
<td>47.02</td>
<td>50.63</td>
</tr>
<tr>
<td>Stroke (%)</td>
<td>Yes</td>
<td>5.98</td>
<td>2.03</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3.49</td>
<td>5.59</td>
</tr>
<tr>
<td>Smoke Status (%)</td>
<td>Never</td>
<td>51.72</td>
<td>55.74</td>
</tr>
<tr>
<td></td>
<td>Former</td>
<td>25.16</td>
<td>24.93</td>
</tr>
<tr>
<td></td>
<td>Current</td>
<td>22.18</td>
<td>19.19</td>
</tr>
<tr>
<td>Alcohol Status (%)</td>
<td>Never</td>
<td>22.19</td>
<td>21.78</td>
</tr>
<tr>
<td></td>
<td>Former</td>
<td>39.69</td>
<td>34.17</td>
</tr>
<tr>
<td></td>
<td>Current</td>
<td>37.97</td>
<td>43.93</td>
</tr>
<tr>
<td>BMI</td>
<td></td>
<td>30.15 ± 5.67</td>
<td>30.08 ± 5.57</td>
</tr>
<tr>
<td></td>
<td></td>
<td>30.17 ± 5.63</td>
<td>30.52 ± 5.85</td>
</tr>
</tbody>
</table>

Table 3: Performance with different number of trees.

<table>
<thead>
<tr>
<th>Number of Trees</th>
<th>Accuracy (Training)</th>
<th>AUROC (Training)</th>
<th>Accuracy (Testing)</th>
<th>AUROC (Testing)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.9050</td>
<td>0.9275</td>
<td>0.6026</td>
<td>0.6010</td>
</tr>
<tr>
<td>50</td>
<td>0.9493</td>
<td>0.9492</td>
<td>0.6731</td>
<td>0.6106</td>
</tr>
<tr>
<td>100</td>
<td>0.9493</td>
<td>0.9492</td>
<td>0.6731</td>
<td>0.6106</td>
</tr>
<tr>
<td>500</td>
<td>0.9544</td>
<td>0.9520</td>
<td>0.6795</td>
<td>0.6154</td>
</tr>
<tr>
<td>1,000</td>
<td>0.9469</td>
<td>0.9497</td>
<td>0.6923</td>
<td>0.6346</td>
</tr>
<tr>
<td>2,000</td>
<td>0.9465</td>
<td>0.9482</td>
<td>0.6975</td>
<td>0.6202</td>
</tr>
</tbody>
</table>

Figure 3: The top 20 important sleep-related variables for different genders: (a) Female-specific and (b) Male-specific.
Table 4: The feature importance (FI) of 47 sleep-related variable candidates.

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>FI</th>
<th>Variable Name</th>
<th>FI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Standard Deviation Heart Rate</td>
<td>1.0000</td>
<td>25. Apnea/Hypopnea Events (1% desat, non-supine)</td>
<td>0.688</td>
</tr>
<tr>
<td>2. Maximum Heart Rate</td>
<td>0.9672</td>
<td>26. Apnea/Hypopnea Index (3% desat)</td>
<td>0.6851</td>
</tr>
<tr>
<td>3. Apnea/Hypopnea Index (all desats)</td>
<td>0.8943</td>
<td>27. Doze Off Situations: Lying down to rest in afternoon</td>
<td>0.6823</td>
</tr>
<tr>
<td>4. Baseline SpO2</td>
<td>0.8929</td>
<td>28. Total Time spent in Apnea/Hypopnea event (3% desat, supine) (seconds)</td>
<td>0.6812</td>
</tr>
<tr>
<td>5. Total Time in Bed (hours)</td>
<td>0.8777</td>
<td>29. Apnea/Hypopnea Events (1% desat, supine)</td>
<td>0.6777</td>
</tr>
<tr>
<td>6. Standard Deviation SpO2</td>
<td>0.8601</td>
<td>30. Wake up earlier than you plan</td>
<td>0.6775</td>
</tr>
<tr>
<td>7. Minimum SpO2</td>
<td>0.8589</td>
<td>31. Total Time spent in Apnea/Hypopnea event (3% desat, non-supine) (seconds)</td>
<td>0.6715</td>
</tr>
<tr>
<td>8. Apnea/Hypopnea Events (all desats)</td>
<td>0.8224</td>
<td>32. Apnea/Hypopnea Events (1% desat)</td>
<td>0.6701</td>
</tr>
<tr>
<td>9. Apnea/Hypopnea Index (1% desat, non-supine)</td>
<td>0.8136</td>
<td>33. Apnea/Hypopnea Events (3% desat, supine)</td>
<td>0.6344</td>
</tr>
<tr>
<td>10. Mean SpO2</td>
<td>0.8035</td>
<td>34. Apnea/Hypopnea Events (3% desat, non-supine)</td>
<td>0.6176</td>
</tr>
<tr>
<td>11. Percent Time SpO2 &lt; 95</td>
<td>0.7986</td>
<td>35. Weekday sleep duration (hours)</td>
<td>0.6400</td>
</tr>
<tr>
<td>12. Total Time spent in Apnea/Hypopnea event (all desats) (hours)</td>
<td>0.7922</td>
<td>36. Apnea/Hypopnea Events (3% desat)</td>
<td>0.6213</td>
</tr>
<tr>
<td>13. Average sleep duration (hours)</td>
<td>0.7811</td>
<td>37. Apnea/Hypopnea Events (3% desat, non-supine)</td>
<td>0.6176</td>
</tr>
<tr>
<td>14. How Often Snore</td>
<td>0.7748</td>
<td>38. Nap for 5 min or more during a usual week</td>
<td>0.6032</td>
</tr>
<tr>
<td>15. Apnea/Hypopnea Index (1% desat)</td>
<td>0.7738</td>
<td>39. Take sleeping pills</td>
<td>0.4506</td>
</tr>
<tr>
<td>16. Typical night’s sleep in past 4 weeks</td>
<td>0.7441</td>
<td>40. Doze Off Situations: Sitting and talking</td>
<td>0.3951</td>
</tr>
<tr>
<td>17. Total Time spent in Apnea/Hypopnea event (1% desat, non-supine) (seconds)</td>
<td>0.7171</td>
<td>41. Doze Off Situations: At the dinner table</td>
<td>0.3624</td>
</tr>
<tr>
<td>18. Sleep difficulties make irritable (SLEA9)</td>
<td>0.7149</td>
<td>42. Overall Study Quality</td>
<td>0.3575</td>
</tr>
<tr>
<td>19. Average time spent in Apnea/Hypopnea event (1% desat, non-supine) (seconds)</td>
<td>0.7035</td>
<td>43. Discomfort in legs worse when at rest</td>
<td>0.3515</td>
</tr>
<tr>
<td>20. Apnea/Hypopnea Index (3% desat, supine)</td>
<td>0.7038</td>
<td>44. Discomfort in legs worse later in day or at night</td>
<td>0.3499</td>
</tr>
<tr>
<td>21. Total Time spent in Apnea/Hypopnea event (1% desat, supine) (seconds)</td>
<td>0.6992</td>
<td>45. Sensation in legs discomfort</td>
<td>0.3494</td>
</tr>
<tr>
<td>22. Trouble getting back to sleep</td>
<td>0.6978</td>
<td>46. Doze Off Situations: While driving</td>
<td>0.3331</td>
</tr>
<tr>
<td>23. Feel Sleepy during the Day</td>
<td>0.6939</td>
<td>47. Need to relieve discomfort in legs</td>
<td>0.3277</td>
</tr>
<tr>
<td>24. Apnea/Hypopnea Index (1% desat, supine)</td>
<td>0.6884</td>
<td>48. Discomfort in legs</td>
<td>0.3224</td>
</tr>
</tbody>
</table>

5 Discussion

5.1 The Hispanic/Latino population

With the increase of diverse populations in dementia research and clinical trials, the risk of leaving behind racial and ethnic minorities is heightened, and improving the understanding of ADRD among ethnic minorities is essential to facilitate diverse representation in prevention and treatment research. Hispanic/Latino covers more than 20 ethnic ancestries, with substantial diversity in ancestry, sociodemographic and cultural characteristics, immigration, and geographic distribution in the U.S.

Although Latinos have sociodemographic disadvantages (e.g., lagging behind Whites and African Americans on average in education), high rates of diabetes, and the lowest healthcare insurance coverage rates of any major ethnic/racial group in the U.S., the average life expectancy of Latinos is four years higher than that of Whites. With long life expectancy, high morbidity, low education, and limited healthcare access, the Hispanic/Latino population faces major risks for age-related disorders, specifically ADRD. In a recent study, Perales-Puchalt et al. reported that Latinos are at higher risk of mild cognitive impairment. Our results in Table 1 indicate that, in addition to well-known factors (e.g., age and education status), sleep-related factors may be associated with cognitive function. Thus, it is important and meaningful to understand whether and how different sleep-related factors may be associated with cognitive function in the Hispanic/Latino population.

5.2 Machine learning based approach

We developed a machine learning approach to find potential sleep-related factors for general cognitive function in a Hispanics/Latinos cohort. In our analytical framework, we used the first RF for covariate analysis and estimation of the importance of different variables. The covariates identified by RF, such as age, education status, yearly household income, gender, and anxiety inventory, were consistent with the confounders in the existing studies, which proved the validity of our method. As shown in Table 2, the SSM results demonstrated that the distributions of each selected covariate in LCF and NC group are similar; for example, before matching, the percentage of yearly household income with “$40,001 - $50,000” in NC group was twice as much as in LCF group (6.39% vs. 3.13%, respectively), and it was much more balanced after matching (2.76% vs. 3.59%). After matching, we used bivariate analysis to select the sleep-related variables, which were considered as candidates (with p-value < 0.1). It reduced the input data.
dimension (more than 50%) and the cost in computation time for learning decision trees in the second RF. As a result, RF performance was improved using selected candidates compared to using all sleep-related variables.

Parameter selection is an important problem in classification as changing parameters may significantly impact the performance of the classifier. We performed several experiments with different settings of the number of trees, and the performance only had slight changes, indicating that random forest is not sensitive to its parameters (i.e., high stability). At the end, we selected the number of trees (N=1,000) for further analysis.

To perform cross-model validation, we compared the importance ranking of sleep-related variables using different models, including six random forest models with the different number of trees and XGBoost\(^3\), which is another widely used machine learning algorithm to obtain feature importance after training. Table 5 shows the comparison results. The ranking of top 20 sleep-related variables in Table 4 obtained from the random forest with 1000 trees (RF-1000) was used as the benchmark for the comparison (see labels A-T in Table 5). As can be seen, although the rankings are slightly different, most of these variables (over 75%) are also in the top 20 with other RF models, indicating that the parameters do not have a significant effect on the results. For the XGBoost model, 60% of the top 20 sleep-related variables overlap with those obtained by the benchmark RF model.

### Table 5: Cross-model ranking comparison of the top 20 sleep-related factors.

| Model   | A     | B     | C     | D     | E     | F     | G     | H     | I     | J     | K     | L     | M     | N     | O     | P     | Q     | R     | S     | T     |
|---------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| RF-1000| 3     | 2     | 2     | 3     | 5     | 7     | 7     | 8     | 9     | 10    | 11    | 11    | 12    | 13    | 14    | 15    | 16    | 17    | 18    | 19    | 20    |
| RF-10  | 4     | 4     | 1     | 13    | 6     | 9     | 8     | 10    | 14    | 11    | 17    | 17    | 17    | 9     | 9     | 13    | 12    | 12    | 8     | 15    | 18    |
| RF-50  | 2     | 5     | 3     | 6     | 4     | 7     | 13    | 10    | 14    | 9     | 12    | 8     | 15    | 18    | 11    | 11    | 11    | 12    | 8     | 15    | 18    |
| RF-100 | 3     | 2     | 1     | 6     | 5     | 4     | 8     | 16    | 11    | 10    | 7     | 14    | 13    | 9     | 12    | 18    | 20    |       |       |       |       |
| RF-500 | 1     | 2     | 3     | 4     | 6     | 5     | 7     | 10    | 8     | 12    | 9     | 11    | 15    | 14    | 13    | 17    | 16    | 19    | 18    | 20    |
| RF-2000| 1     | 2     | 4     | 3     | 6     | 5     | 7     | 9     | 8     | 12    | 10    | 11    | 15    | 13    | 14    | 17    | 16    | 19    | 18    | 20    |
| XGBoost|       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |

#### 5.3 Consistency with existing research findings

The results in Table 4 show that several variables play important roles in classifying LCF cases and NCs, which were also consistent with existing research findings, such as: (1) heart rate – heart rate variability was reported as a promising early biomarker of cognitive impairment in populations without dementia or stroke\(^4\), and cardiovascular risk is also a risk for cognitive function in middle-age among Hispanics/Latinos\(^5\); (2) sleep duration – the previous study with HCHS/SOL demonstrated that people with long sleep duration (> 9 hours per night) had a worse cognitive function, and people with intermediate sleep duration (6 to 8 hours per night) had the best cognitive function\(^6\); (3) trouble falling asleep – longer sleep-onset latency has been reported to affect global cognitive function through decreased slow-wave sleep and was associated with declines in verbal learning and memory in the previous study with HCHS/SOL\(^7\); and (4) apnea/hypopnea index – several existing research demonstrated that severe obstructive sleep apnea patients may have impaired cognitive function\(^8\), the treatment of sleep apnea syndrome slows cognitive decline\(^9\), and Ramos et al. reported that obstructive sleep apnea was associated with poor cognitive function in a cohort of Hispanic/Latino women in the U.S.\(^10\). The consistency with existing research findings in part validates the effectiveness of our DRF method. Also, our results were consistent across gender-specific analyses (female vs. male).

#### 5.4 Potential sleep-related factors

We categorized the 47 sleep-related variable candidates in to 7 categories: (1) \textit{SpO2 measures}, including “Baseline SpO2”, “Minimum SpO2”, “Mean SpO2”, “Standard Deviation SpO2”, and “Percent Time SpO2 < 95”; (2) \textit{Sleep durations}, including “Total Time in Bed (hours)”, “Average sleep duration (hours)”, and “Weekday sleep duration (hours)”; (3) \textit{Trouble to sleep}, including “Trouble falling asleep”, “Wake up several times at night”, “Wake up earlier than you plan”, “Take sleeping pills”, “Sleep difficulties make irritable (SLEA9)”, and “Trouble getting back to sleep”; (4) \textit{Feel sleepy}, including “Feel Sleepy during the Day”, “Typical night’s sleep in past 4 weeks”, “Doze Off Situations: Lying down to rest in afternoon”, “Doze Off Situations: Sitting and talking”, “Doze Off Situations: At the dinner table”, “Doze Off Situations: While driving”, and “Nap for 5 min or more during a usual week”; (5) \textit{Legs discomfort}, including “Sensation in legs discomfort”, “Need to relieve discomfort in legs”, “Discomfort in legs worse when at rest”, and “Discomfort in legs worse later in day or at night”; (6) \textit{Heart Rate}, including “Maximum Heart Rate” and “Standard Deviation Heart Rate”; and (7) \textit{Apnea/Hypopnea measures}, which includes the rest of variables, such as
We calculated and ranked the average feature importance for these 7 categories: Heart Rate (0.98), SpO2 measures (0.84), Sleep durations (0.76), Apnea/Hypopnea measures (0.70), Trouble to sleep (0.65), Feel sleepy (0.54), and Legs discomfort (0.34). The consistency with existing research findings indicate that our method can be used to explore additional potential sleep-related factors that may be associated with cognitive function. As shown in Table 4, sleep-related variables, such as “Mean SpO2” in SpO2 measures, “Wake up several times at night” and “Trouble getting back to sleep” in Trouble to sleep, and “Feel Sleepy during the Day” in Feel sleepy, may be associated with cognitive function for Hispanics/Latinos population.

5.5 General applicability

Although our DRF approach was developed to identify sleep-related factors for cognitive function, its framework (Figure 1) has been designed and implemented to be generally applicable for factor exploration of other diseases and cohorts. For a new disease/cohorts, the DRF approach can be readily used once the data is provided according to the needs of the disease/cohorts.

5.6 Limitations and future work

In this paper, we mainly focused on estimating the importance of various sleep-related variables for discriminating LCF cases and NCs. The relationship/association among these variables has not been studied. In future work, we plan to explore the correlations among these variables. In addition, finding a representative variable for each sleep characteristic instead of using multiple variables (such as “Mean SpO2”, “Minimum SpO2”, and “Standard Deviation SpO2”) is still a challenge and needs further investigation.

6 Conclusion

In this paper, we analyzed sleep-related factors to differentiate a lower cognitive function group from a normal control group in a Hispanics/Latinos cohort using a novel dual random forest approach. The random forest-based analysis enables the estimation of the importance of sleep-related variables for cognitive function. Our analysis obtained several specific sleep-related factors that may be associated with cognitive function in the Hispanics/Latinos cohort, such as heart rate, sleep duration, trouble falling asleep, and apnea/hypopnea index, which are consistent with existing research findings. Our dual random forest approach is effective in exploring the association between sleep-related factors and cognitive function.

Acknowledgment

This work was supported by the National Institute on Aging of the National Institutes of Health under Award Number R21AG068994. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Imputing Longitudinal Growth Data in International Pediatric Studies: Does CDC Reference Suffice?

Zhiguo Li, PhD1, Jorma Toppari, MD, PhD2, Markus Lundgren, MD, PhD3, Brigitte I. Frohnert, MD, PhD4, Peter Achenbach, MD5, Riitta Veijola, MD, PhD6, and Vibha Anand, PhD7 for the T1DI study group
1Center for Computational Health IBM Research, NY, NY, 2Institute of Biomedicine and Population Health Research Centre, University of Turku and Department of Pediatrics, Turku University Hospital, Turku, Finland, 3Department of Clinical Sciences Malmö, Lund University/CRC, Skåne University Hospital, Malmö, Sweden, 4Barbara Davis Center for Diabetes, University of Colorado, Denver, CO, USA, 5Institute of Diabetes Research, Helmholtz Zentrum München, German Research Center for Environmental Health, Munich-Neuherberg, Germany, 6Department of Pediatrics, PEDEGO Research Unit, University of Oulu and Oulu University Hospital, Oulu, Finland, 7Center for Computational Health IBM Research, Cambridge, MA

Abstract

This study investigates a missing value imputation approach for longitudinal growth data in pediatric studies from multiple countries. We analyzed a combined cohort from five natural history studies of type 1 diabetes (T1D) in the US and EU with longitudinal growth measurements for 23,201 subjects. We developed a multiple imputation methodology using LMS parameters of CDC reference data. We measured imputation errors on both combined and individual cohorts using mean absolute percentage error (MAPE) and normalized root-mean-square error (NRMSE). Our results show low imputation errors using CDC reference. Overall height imputation errors were lower than for weight. The largest MAPE for weight and height among all age groups was 4.8% and 1.7%, respectively. When comparing performance between CDC reference and country-specific growth charts, we found no significant differences for height (CDC vs. German: p=0.993, CDC vs. Swedish: p=0.368) and for weight (CDC vs. Swedish: p=0.513) for all ages.

Introduction

The problem of missing data commonly exists in observational studies.1 We explored imputation methodology to address missing values in childhood growth data from prospective natural history studies of type 1 diabetes (T1D), including two in the US (DAISY2, DEW-IT3) and three in the EU (BABYDIAB4, DiPiS5, DIPP6). The primary aim of these studies was to study development of islet autoimmunity (IA) and progression to type 1 diabetes (T1D) in infants and young children with high genetic or familial risk. Study subjects were followed for development of IA for a period of up to 15 years or until diagnosis of T1D, whichever came first. Data collection efforts mainly focused on immunological (islet autoantibodies) and metabolic markers (blood glucose levels) that are implicated in development of T1D.

In a collaboration with JDRF and their academic partners, the type 1 data intelligence (T1DI) study group combined and harmonized data from these large observational studies into the T1DI cohort. The cohort has over 24,000 subjects and the study group is currently evaluating various outcomes of interest from this large dataset, including the potential effect of childhood growth on risk of IA and development of T1D. In practice, childhood growth is often assessed via measurement of height and weight at periodic intervals (e.g. 6m, 1y, 2y visit) during pediatric office visits. During study follow, a subject may not have growth assessment at every visit for blood sampling for autoantibody or blood glucose measurement. In order to test interesting hypotheses, such as the effect of childhood growth on IA development, we need to link available growth data from patient charts with research data measured at more frequent intervals (i.e. every 3 to 9 months in these studies). However, this linking, which may be fuzzy based on nearest age visit or similar methods) often creates holes (sparsity) in the longitudinal research data making them difficult or impossible to analyze unless meaningful imputations are performed for growth variables.

In the T1DI cohort, missing data exist in weight and height measurements of children participating in follow-up. For meaningful inclusion of these variables in association analyses or predictive models (such as for IA development and T1D onset), data need to be imputed to coincide with timing of research visits. Furthermore, on the
basis of the observed/imputed body weight and height, other growth measurements such as the Body Mass Index (BMI) or other derived features such as percentiles and velocities of weight, height, and BMI may be computed which can then be also included in downstream models. Of note imputation of growth data for the T1DI cohort is hindered by a lack of a common reference for childhood growth data. The T1DI cohort growth data are from disparate geographic areas with variable demographic compositions and applicable population growth references may vary. For example, in the US, CDC growth charts\textsuperscript{7} are used widely and are parameterized for childhood and adolescent growth from age 2 to 20 years in intervals of 6 months\textsuperscript{8}. In many EU countries, country-specific growth charts or references exist.\textsuperscript{9,10} However, these references are either not easily available or parameterized for similar use. Thus, in this study we hypothesized that the childhood growth in the T1DI cohort could be imputed using the CDC LMS parameters. We further hypothesized that using the CDC reference for imputation would produce overall low error rates. This approach could provide several other advantages in large international studies such as use of one common reference, model and fit. In this work we test our hypotheses using a novel multiple imputation (MI) methodology as described below.

Regardless of the reference data used, missing data should be handled and analyzed with caution depending on variable types and their mechanism of missingness. The missingness mechanism can be broadly classified into three types according to the characteristics with regards to randomness\textsuperscript{11}: missing not at random (MNAR), missing at random (MAR), and missing completely at random (MCAR). Significant interactions between observed variables are an indicator of MNAR data. In general, missing value imputation for MNAR is not recommended, and hence few algorithms exist for this type of data. In contrast, a wide variety of methods are available for handling missing values in MAR and MCAR types. In practice, selecting the most appropriate method depends on the problem to be addressed and the available data. Single imputation methods simply replace missing data points with a single fixed value (e.g., zero, the mean, median, or most frequent value) by assuming that the data are MCAR. These methods could reduce variability of data and generate biased estimates of error variances.\textsuperscript{12} Advanced methods such as model-based methods impute the missing values using a predictive distribution that models the underlying data missingness mechanism. The most commonly used model-based imputation methods include multiple imputation (MI), full information maximum likelihood (FIML) and maximum likelihood expectation-maximization (EM) imputation. In particular, MI assumes MAR and has been widely applied in analysis of clinical data.\textsuperscript{13} Some popular algorithms in this category include nonparametric missing value imputation using random forest\textsuperscript{14}, K-nearest neighbors (KNN), multiple imputation by chained equations (MICE)\textsuperscript{15}, and Amelia.\textsuperscript{16} More recently, more sophisticated models have been used for missing value imputation including multilayer perceptron, self-organizing maps as the predictive models. These have been applied to estimation of missing data for applications such as breast cancer diagnosis and cardiovascular data for clinical trials.\textsuperscript{17,18} In our study, we use a general MI method and apply it to missing values of growth data in pediatric studies.

**Methods**

To describe characteristics of data missingness, we calculated missing value ratio. CDC growth charts and LMS parameters as well as country-specific growth charts are used for missing value imputations. We describe these sources below followed by our particular implementation of the MI algorithm, which considers uncertainty of the missing data and combines multiple sets of plausible imputed values.

**Data Missingness:** Missingness for height and weight were separately assessed for each subject in every dataset. We define missing value ratio as the number of rows (visits) where measurement (for height or weight separately) was missing divided by the total number of rows for that subject in the longitudinal dataset. For example, if the missing value ratio of height is 0.1, then 10% of longitudinal measurements of height were missing for that subject.

**CDC Growth Charts and LMS Parameters:** In the United States, the CDC growth charts\textsuperscript{19} are widely used to monitor growth patterns of children and adolescents during clinical care. CDC charts describe values for height or weight at different percentiles by gender and age in months (from birth to 20-years in increment of 1 month). The smoothed percentile curves can be reproduced using LMS parameters, i.e. median (M), generalized coefficient of variation (S), and the power in the Box-Cox transformation (L). These parameters were calculated based on the Box-Cox transformations to normality. From these parameters, a measurement value can be calculated if the percentile is known, and vice versa. CDC growth charts span the age range of 0 – 36 months and 2 – 20 years. The LMS parameters, for 3rd, 5th, 10th, 25th, 50th, 75th, 90th, 95th, and 97th percentiles for weight, height and BMI are provided. For finer age intervals, interpolation could be used to obtain the LMS values. In this study, we use the
CDC reference to impute missing values in the body weight and height using the LMS parameters. We evaluate these values against the country-specific growth charts.

**Country-specific growth charts:** These were obtained from our collaborators for Swedish (DiPiS) and German (BABYDIAB) cohorts. There was no similar reference chart available for the Finnish cohort to our team. We used country-specific charts to impute missing values and compared these with CDC growth charts. The Swedish growth charts cover the age range of 0 – 2191 days (72 months) and 61-228 months. It provides the values of standard deviations (-5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5) for weight, height, and BMI. These growth charts were used for imputing missing values in DiPiS. Note that the LMS parameters are not provided in these charts. The German growth charts cover the age range of 0 – 216 months. They list the LMS parameters and the percentiles (3rd, 10th, 25th, 50th, 75th, 90th, and 97th) for weight and height. They also provide the LMS parameters and the percentiles (3rd, 10th, 50th, 90th, and 97th percentile) for BMI. These charts were used for imputing missing values in BABYDIAB.

Of note, CDC growth charts cover measurements for up to 20-years of age, the German growth charts up to 18 years, and the Swedish charts up to 19 years. Thus, to address measurements until age of 20, extrapolation was used for the latter two cohorts. In this study, we imputed the missing values in the body height and weight using CDC growth charts for all five (US and EU) data sets.

**Missing Value Imputation method:** We use the growth charts with LMS parameters for missing value imputations and in that respect, it is a model-based method. Additionally, we apply multiple imputations to derive the missing value, i.e. “multiply impute” missing data in a single cross-section from an available time series of data. The flow chart is illustrated in Figure 1. The work was implemented using R (v3.6.3). The inputs of this algorithm include sex and age of a subject; the outputs are the imputed values and the corresponding percentiles for height/weight. This missing value imputation procedure consists of the following steps: For a growth measure (height or weight),

- **Step I:** For an individual subject, locate the ages at which the measurement values were missing and find the sex. Then identify the most proximal previous and subsequent ages with observed values. For each of the ages with missing values, repeat the following steps II-IV:

  - **Step II:** For the previous age with known measurement value, use growth chart to compute the corresponding LMS values for that age through interpolation (if needed). Using the observed measurement and interpolated LMS values, compute its percentile (the percentile of the previous measurement) using the LMS formula. Do the same for the subsequent visit with a measured value to compute its percentile.

  - **Step III:** if both the previous and the subsequent percentiles are obtained, take the average of these two as the percentile for the age at which we want to impute the missing value; otherwise, just use the available percentile (either the previous percentile or subsequent percentile).

  - **Step IV:** for the age with a missing value, use the growth charts to obtain the LMS values through interpolation of the interval covering that age. Using the percentile value calculated in Step III and the interpolated LMS values to determine the imputed value, which is used to replace the missing measurement at that age.

---

**Figure 1. Flow chart for missing value imputation using the growth charts**
Note that in Step III, in case that only a single percentile is available, extrapolation is used to impute the missing values and could cause larger imputation errors since it assumes that the growth percentile remains unchanged regardless of the length of time interval.

Performance measures: To assess the validity of missing value imputation method and compare the performance of the missing value imputation using country specific growth charts, we calculated imputation errors as follows. First, we randomly selected 20% of observed measurements and then removed their values creating a knock-out dataset. We then applied the missing value imputation method to the knock-out dataset and imputed the “missing” values. Finally, the imputation errors were computed. The same procedure was implemented using CDC or country-specific references. We used the following metrics for the imputation errors—mean absolute percentage error (MAPE) and normalized root-mean-square error (NRMSE). MAPE is defined as

\[
MAPE = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{y_{i}^{imp} - y_{i}^{act}}{y_{i}^{act}} \right|
\]

where \(y_{i}^{imp}\) and \(y_{i}^{act}\) are imputed and actual (observed) measurement values, respectively, for \(i = 1, 2, \ldots, n\) and \(n\) is the total number of imputed values. There is some variation in NRMSE definitions; however, the most common one we used in this study as follows,

\[
NRMSE = \frac{\frac{1}{n} \sum_{i=1}^{n} \left( y_{i}^{imp} - y_{i}^{act} \right)^{2}}{\frac{1}{n} \sum_{i=1}^{n} y_{i}^{act}}^{1/2}
\]

where the denominator is the average of actual measurements. From the above definitions, we can see that MAPE is more sensitive than NRMSE to the errors when the actual values are small, while NRMSE puts more weight on larger deviations than MAPE.

We calculated mean and standard deviation of MAPE or NRMSE values using 100 repetitions of knock-out and imputations and computed p-values for comparison with country specific references. All significance was tested at \(P < 0.05\) using the Student’s \(t\) test.

Results

In total there are 23,201 subjects for whom growth data were available (Table 1). Overall DIPP is the largest dataset with a frequent follow up compared to the overall T1DI cohort, and consequently had the smallest missing value ratios for both height and weight in our study. However, the two US datasets (DAISY and DEW-IT) had lower missing value ratios (for both height and weight) when compared to the other two EU sites (BABYDIAB and DiPiS). These differences are primarily due to differences in the individual study protocols and the frequency of follow up within them.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Number of Subjects</th>
<th>Height Missing Value Ratio Median [IQR]</th>
<th>Weight Missing Value Ratio Median [IQR]</th>
</tr>
</thead>
<tbody>
<tr>
<td>BABYDIAB</td>
<td>2340</td>
<td>0.4 [0.25-0.50]</td>
<td>0.38 [0.25-0.50]</td>
</tr>
<tr>
<td>DAISY</td>
<td>2148</td>
<td>0.18 [0.09-0.50]</td>
<td>0.0 [0.0-0.06]</td>
</tr>
<tr>
<td>DEW-IT</td>
<td>2830</td>
<td>0.13 [0.0-0.73]</td>
<td>0.0 [0.0-0.50]</td>
</tr>
<tr>
<td>DiPiS</td>
<td>4227</td>
<td>0.67 [0.43-0.73]</td>
<td>0.64 [0.43-0.73]</td>
</tr>
<tr>
<td>DIPP</td>
<td>11656</td>
<td>0.0 [0.0-0.05]</td>
<td>0.0 [0.0-0.05]</td>
</tr>
</tbody>
</table>

**Imputations in combined T1DI cohort:** Figure 2 illustrates the MAPE and NRMSE values for missing value imputations in weight and height for the combined dataset (from 5 data sources) using the CDC growth charts. The MAPE values are plotted by the age group (6 groups from birth to 20-years as covered by CDC) as labeled on the y-axis. For the 0-2 year age group, the largest MAPE value for weight is 4.8% and the largest error for height is 1.7%.
The NRMSE values of the imputation errors are illustrated in Figure 2(b). The largest value is 6.0% for weight in age group 18-20 years and 2.1% for height in 0-2 years. From Figure 2(a) and 2(b), we find that the imputation errors for height are overall smaller than those for weight (in terms of both MAPE and NRMSE). In addition, for the same growth measure (height or weight), the MAPE values are smaller than the NRMSE values because NRMSE puts more weight on larger absolute measurement errors.

Figure 2. Imputation errors in weight and height for TIDI cohort using CDC (a) MAPE; (b) NRMSE

Figure 3-1. Imputation errors in weight and height for DAISY using CDC (a) MAPE; (b) NRMSE

Figure 3-2. Imputation errors in weight and height for DEW-IT using CDC (a) MAPE; (b) NRMSE

Figure 3-3. Imputation errors in weight and height for DIPP (a) MAPE; (b) NRMSE
**Imputations in US datasets:**

The MAPE and NRMSE values for two US data sources (DAISY and DEW-IT) are shown in Figure 3-1 (a and b) and Figure 3-2 (a and b) respectively. The imputation was implemented using CDC as reference. For DAISY, the largest MAPE values of weight are 5.1% for 0-2 years and 1.9% for height in the same age range. The largest NRMSE values for the same age group are 6.5% for weight and 2.3% for height. When compared to DAISY, the DEW-IT dataset has larger imputation errors in both height and weight. In terms of MAPE, the largest errors are 6.8% in weight and 3.3% in height for 0-2 years. The largest NRMSE values are 10.2% in weight and 4.6% in height for 8-12 years.

**EU datasets and comparison to country-specific growth charts:**

As the Finnish growth charts were not available for comparison, only CDC charts were used for imputing missing values in DIPP. The imputation errors in DIPP for height and weight can be found in Figure 3-3 (a) MAPE and (b) NRMSE. The largest MAPE values for weight using CDC growth charts are 4.1%, and 1.9% for height. The largest NRMSE values are 5.0% and 1.9% for weight and height, respectively.

For the other 2 EU datasets, we compared performance of imputation method using country specific growth charts. For BABYDIAB, we compared imputation method using the CDC growth charts and the German growth charts. For simplicity, we only plotted the MAPE values here on Figure 4-1 (a - height and b - weight). Since the German charts cover age up to 18 years, only the results from birth to 18 years are shown here. The largest MAPE values for weight in BABYDIAB (Figure 4-1) using CDC and German growth charts are 6.8% and 6.7% respectively, and the largest MAPE for height are 2.4 % and 2.3%, respectively. The Student’s test results computed for all ages together show no significant difference in the MAPE values of height using CDC ($p =0.993$), but the mean MAPE for weight using CDC is greater than that using the German charts ($p < 0.001$). However, the difference between them was fractional (6.02% vs. 5.81%).

![Figure 4-1. MAPE of the imputed values in weight and height for BABYDIAB (a) height; (b) weight](image1)

![Figure 4-2 MAPE of the imputed values in weight and height for DiPiS (a) height; (b) weight](image2)

The MAPE results of DiPiS are shown in Figure 4-2 (a- height and b-weight) for which both CDC charts and Swedish charts were used. The age range of DiPiS is 0 – 12.99 years, therefore the largest age group is also 12-18 years. Due to the large missing value ratio, the largest MAPE values for weight in DiPiS using CDC and Swedish growth charts are 16.4% and 15.5% respectively, and the largest MAPE for height are 6.0 % and 5.4%, respectively.
The Student’s t test results computed for all ages together show that there is no significant difference in the MAPE values of height using CDC or Swedish growth charts ($p = 0.37$), and this is also true for weight as well ($p = 0.51$).

**Discussion**

In this study, we developed a missing value imputation method for data from multiple countries using CDC growth charts and a multiple imputation methodology. We assessed imputation performance using the large T1DI cohort and its constituent five real-world data sets. We also compared the CDC-imputed datasets to country-specific growth charts where available. From these experiments, we show that CDC reference may suffice to impute growth measurements in US and international (EU) pediatric studies. We further show that using this method, the imputation errors (as measured by MAPE and NMRSE) are within reasonable errors of margin (<10%) in the T1DI cohort. Furthermore, we found that the errors for imputed height are less than those for imputed weight across all datasets. This can be explained by the fact that height is much more tightly regulated by genes than weight which is strongly influenced by environment and eating behavior. Thus, height can be easily predicted on the basis of surrounding growth pattern, whereas weight can vary much over time.

There are a few country-specific findings. The imputation errors in BABYDIAB (Germany) and DAISY (U.S.) were much smaller than those in DIPIS (Sweden). These may be due to two reasons, overall low missing value ratio in the DAISY and BABYDIAB data sets and well calibrated CDC and German growth charts. However, there is no significant difference between the error results using the CDC reference vs. the country-specific charts. The imputation errors are higher in DiPIS perhaps because of larger missing value ratios compared to the other data sets (Table 1) so that the average of the previous and subsequent observed percentiles are not a good estimate for the percentile at the age with missing growth measures.

There are several advantages of our approach. First, one common CDC growth reference for imputation in all datasets may facilitate imputations where standards don’t exist or are unavailable, such as for the Finnish dataset. As the CDC reference comes from a large, diverse US sample population and has been studied extensively, it may also be the best guesstimate where growth patterns are unknown (or less studied) and almost certainly change by age groups (e.g. from early childhood to tween and adolescent age groups). Furthermore, using imputed growth measurements based on a reference model, one can derive many other features such as annual growth velocity and percentiles for downstream analyses. A second advantage is that the CDC reference covers a wide age range (0 to 20 years); therefore, one can use it to impute the missing values over the entire growth trajectory. This in comparison to a best fit approach offers significant advantages when rate of change (or slope) in longitudinal data are unknown and where a change detection by age is needed. In these latter scenarios multiple models may be needed to assess different segments. Third, our approach is also useful when the task at hand is to perform an association study (limiting data to a landmark age for prediction in future). In our use case of studying growth features and their potential contribution to development of islet autoimmunity, we found that imputations were reasonable. Finally, as the approach relies on the underlying model, our methods may be useful for other kinds of clinical data, for example for imputing missing blood pressure or glucose measurements.

As with all such studies, our approach also has some limitations. First, we base our results on data collected in the past (in many instances these studies started before 2000), i.e. before the CDC reference was widely available. However, this may have only made our results stronger in showing there is no significant difference between country-specific or CDC reference-based imputations over time. Second, it may not be the best approach to use when one needs to analyze outcomes based on observed growth to a time point (i.e. where we cannot avail ourselves of future information). In those situations, another model-based imputation method (e.g. using linear or polynomial multi-variable regression) may be a better fit. We believe the strength of our multiple imputation approach lies in where references such as the CDC are available for characterizing longitudinal pediatric growth trajectory. In addition, since the approach assumes MAR for the missing data, departures from this assumption should be assessed through sensitivity analyses.

**Conclusion**

Childhood growth variables measured during follow-up of children and adolescents up can be imputed using multiple imputation method and CDC reference parameters even when data represents geographically disparate US
and EU pediatric sites. Advantage of this approach include being able to derive many other features, such as annual growth velocity and growth percentiles; these features may be useful for downstream analyses.

Acknowledgments


References


17. Rahman MM, Davis DN. Fuzzy Unordered Rules Induction Algorithm Used as Missing Value Imputation Methods for K-Mean Clustering on Real Cardiovascular Data.


Reducing Physicians’ Cognitive Load During Chart Review: A Problem-Oriented Summary of the Patient Electronic Record

Jennifer J. Liang, MD, Ching-Huei Tsou, PhD, Bharath Dandala, PhD,
Ananya Poddar, MS, Venkata Joopudi, MS, Diwakar Mahajan, MS, John Prager, PhD,
Pree thi Rag havan, PhD, Michele Payne, BS
IBM T.J. Watson Research Center, Yorktown Heights, NY

Abstract
Overabundance of information within electronic health records (EHRs) has resulted in a need for automated systems to mitigate the cognitive burden on physicians utilizing today’s EHR systems. We present ProSPER, a Problem-oriented Summary of the Patient Electronic Record that displays a patient summary centered around an auto-generated problem list and disease-specific views for chronic conditions. ProSPER was developed using 1,500 longitudinal patient records from two large multi-specialty medical groups in the United States, and leverages multiple natural language processing (NLP) components targeting various fundamental (e.g. syntactic analysis), clinical (e.g. adverse drug event extraction) and summarizing (e.g. problem list generation) tasks. We report evaluation results for each component and discuss how specific components address existing physician challenges in reviewing EHR data. This work demonstrates the need to leverage holistic information in EHRs to build a comprehensive summarization application, and the potential for NLP-based applications to support physicians and improve clinical care.

Introduction
Electronic health records (EHRs) are one of the primary sources of clinician burnout. Many studies have analyzed physician time spent in EHRs with respect to various interactions such as chart review, documentation, orders, and inbox management, with several identifying time spent on reviewing information within the EHR, i.e. chart review, as a major component of physician interaction with the EHR. A descriptive study on ambulatory medical subspecialists and primary care physicians observed that chart review accounted for the highest proportion of active physician EHR time at 33%. Another study analyzed audit logs to measure time requirements for EHR use by ophthalmologists and concluded that most of the time spent in the EHR is on reviewing information.

Several factors contribute to the difficulty in reviewing information in EHRs. Traditionally, the patient medical record was mainly used to document a patient’s medical history and clinical care process to assist physicians in providing informed care. However, the transformation from paper-based to electronic patient records have expanded the scope of documentation to also serve a variety of non-clinical purposes, including administrative, legal, research, and education. This transformation has resulted in lengthy and complex documentation dominated by content not directly relevant to clinical care, such as text meant for billing purposes, quality improvement measures, avoiding malpractice, and documenting compliance. Further exacerbating this problem are issues of note bloat, where use of structured data imports, templates and copy-paste have introduced unnecessary, redundant or erroneous data into clinical notes, worsening the problem of information overload and physician stress. A human factors engineering perspective to the information chaos in primary care identified five information hazards: information overload, information underload, information scatter, information conflict, and erroneous information. These information hazards result in additional workload for physicians, such as spending extra effort to search through charts, asking more questions of the patient to clarify conflicting documentation, or re-working a diagnosis to investigate potentially erroneous information. Together these factors contribute to excess clinician workload and burnout, with adverse outcomes on clinician well-being, patient care, and the health care system.

Previous research has explored several automated summarization and enhanced visualization methods to enable physicians to efficiently navigate and discern complex data within EHRs. Summarization efforts include problem identification from a given set of clinical notes, extraction of clinical concepts or relations, and various ways to display such information along with other information in the EHR. Another research direction has focused on improved data visualization and interactive displays to facilitate more efficient and accurate information retrieval and an overall more user-friendly interaction with the EHR. Evaluation of such systems in real or simulated clinical settings have received positive feedback from physicians and resulted in improved physician performance. A common theme in many of these systems is the idea of a problem-oriented medical record (POMR), and the use
of timelines to gain a longitudinal history of the patient. We rely on these principles to build Problem-oriented Summary of the Patient Electronic Record (ProSPER), an application which extends the idea of a POMR by enhancing it with clinically meaningful insights extracted from unstructured clinical notes, such as adverse drug events (ADEs), and displaying them on a timeline with relevant structured elements of the EHR.

During the past several years, we have developed and validated the usability of several components that contribute to the creation of a comprehensive patient record summarization application. The purpose of this paper is to build on our previously published work and present ProSPER, an artificial intelligence-based system that extracts, analyzes, and summarizes patient clinical information from electronic health records, with the goal of reducing the cognitive burden on physicians due to information chaos within the EHR. We present the overall system architecture and pipeline, describe the underlying components needed to build such a system, and discuss how these components address information chaos within the EHR. ProSPER differs from previously published systems in that it presents (1) an auto-generated problem list, (2) disease-specific dashboards, (3) a stacked timeline view incorporating both structured data and insights extracted from unstructured data to allow reasoning across different data elements, and (4) a textual summary of each clinical encounter for a selected disease of interest.

Materials and Methods

Combining the well-established idea of a problem-oriented medical record with the different workflows using EHRs that rely on electronic dashboard summary screens as well as individual progress notes, we envision a workflow for reviewing patient information prior to a clinic visit as follows.

1. Patient overview: gain an overview of the patient, including the patient’s name, demographics, allergies, and a summary of the active problems, current medications, allergies, vital signs, and laboratory results.
2. Snapshot of a specific problem: focus on a specific disease by surfacing current medications and recent laboratory results and procedures relevant to the specified disease.
3. Longitudinal history of a specific problem: investigate a specific disease through a timeline view of relevant medications, laboratory results, procedures, past encounters and clinical events extracted from clinical notes.

Dataset

Effective natural language processing (NLP) techniques are essential to development of the proposed system. To achieve this, we used a set of 1,500 de-identified longitudinal patient records from two large ambulatory multi-specialty medical groups in the United States. Subsets of this corpus were manually annotated by medical students and practicing physicians to generate ground truth for development of smaller NLP components that are used in building the overall system. All annotations were doubly-annotated and adjudicated to ensure the quality and consistency of the ground truth. Details of the ground truth generation process for each task are provided below.

The dataset for problem list generation consists of 399 randomly selected patient records annotated by over 17 fourth year medical students. Annotators reviewed each patient record, created a problem list for a comprehensive health assessment, normalized identified problems to the clinical observations recordings and encoding (CORE) problem list subset of Systematized Nomenclature of Medicine-Clinical Terms (SNOMED CT) in a one-to-many mapping, and evaluated how closely the SNOMED CT concept represents the problem.

The dataset for relation detection consists of 24,537 unique problem-medicine pairs, 11,318 unique problem-procedure pairs, and 23,402 unique problem-laboratory test pairs created using structured data from 100 randomly selected patient records. Four fourth year medical students reviewed each pair of entities and indicated if a positive association was present, defined as medications that treat or prevent the problem or its symptoms, and laboratory tests or procedures that can be used to screen, prevent, evaluate, diagnose, or manage the problem.

The dataset for extractive summarization consists of 3,453 clinical notes over 762 patients with recent documentation of at least one of five common chronic diseases (hypertension, diabetes mellitus, hyperlipidemia, heart failure, chronic obstructive pulmonary disease) and authored by physicians or nurse practitioners. The corpus was annotated by 12 physicians with emphasis on only annotating the most important insights relevant to physicians’ decision-making.

The ground truth for medication insights, specifically ADE and medication change events, was annotated at the note-level. For ADEs, 602 notes were annotated with medications and their associated adverse events. For medication change events, 38,895 medication mentions were annotated to indicate (1) the presence of a medication change event and (2) the multi-dimensional context (action, temporality, certainty) for identified medication change events.
The dataset for evaluation of semantic find is derived from 10 randomly selected patient records. For each patient record, a physician generated a set of relevant search terms based on the patient’s last progress note and problem list. The final dataset consists of 169 search terms, covering a variety of semantic types (e.g., symptoms, drugs), single and multi-word concepts, different parts of speech, and commonly-accepted medical abbreviations.

**Pipeline and framework**

ProSPER was built using the open-source Apache Unstructured Information Management Architecture (UIMA) library, a software architecture that provides capabilities to construct pipelines consisting of a sequence of components or analysis engines (AE) that inspect the input data, perform analysis, and store resulting annotations. The system consists of several such components that work together to process an input EHR. These components write their output to a shared data structure called a common analysis structure (CAS). The system achieves high performance by using UIMA Asynchronous Scaleout, a set of capabilities supported for achieving scale-out and Distributed UIMA Cluster Computing, a cluster management system providing tooling, management, and scheduling facilities to automate the scale-out of UIMA applications. Figure 1 shows the high-level flow and underlying NLP components of ProSPER. ProSPER’s NLP components, all tailored towards EHR text, can be broadly divided into three levels, described below.

![Figure 1. A high-level architecture of ProSPER.](image-url)
Fundamental and clinical NLP components

Despite the success of current open-domain NLP systems for syntactic and semantic analysis, several studies\textsuperscript{34,35} have shown that substantial effort is needed to adopt existing systems to the clinical domain. Clinical notes are written under considerable time pressure, using a combination of ad-hoc formatting and liberal use of parenthetical expressions, jargon and acronyms to increase the information density\textsuperscript{35}. For example, sentence ends are frequently indicated by layout rather than punctuation, and white space is not always present to indicate token boundaries (eg, “50mg”). To handle such issues, two deep parsing components\textsuperscript{36} were used: a domain-adapted English Slot Grammar parser for core linguistic analyses (tokenization, segmentation, morpho-lexical analysis, syntactic analysis), and a predicate-argument structure (PAS) builder for simplification and abstraction of the ESG parse. These components incorporate domain-specific rules sensitive to low-level features such as punctuation, capitalization, text-wrap properties, and indentation of clinical documentation.

Rule-based, machine learning (ML)-based and lookup-based methods are three major approaches for extracting useful information from natural language texts. Many components in ProSPER use a combination of hand-crafted rules and ML-based methods for their operation. The rule engine in ProSPER uses a pattern language for describing rules on the output of PAS builder. Details about this rule engine are included in Boguraev et al.\textsuperscript{37} Two ML frameworks, feature-based learning and deep learning frameworks, were developed to support concept and sentence classification, sequence labeling, sequential sentence labeling, sentence pair scoring, and relation extraction. The feature-based learning framework provides implementations of traditional ML algorithms such as Support Vector Machines\textsuperscript{38}, Conditional Random Fields (CRF)\textsuperscript{39}, and domain-independent and dependent feature extractors. Similarly, the deep learning framework provides implementations of state-of-the-art deep learning architectures, such as Bidirectional Long Short-Term Memory – Conditional Random Fields\textsuperscript{40} and pre-trained transformer-based models\textsuperscript{41}. These frameworks were tailored to more effectively process clinical notes. They provide the ability to (1) easily integrate feature extractors or architectures, (2) leverage evidence from both structured and unstructured data simultaneously to develop consolidated models, and (3) streamline models or feed predictions from one or more auxiliary tasks to subsequent tasks. ProSPER uses UIMA’s ConceptMapper\textsuperscript{42}, a tool for lookup-based methods, to link text in input documents with entries in a dictionary such as the Unified Medical Language System (UMLS) semantic network\textsuperscript{35}.

Several NLP components that extract useful information from clinical narratives have been developed using these frameworks. Details of these components have been described in our previous studies\textsuperscript{40,42,44–50}. These include (1) fundamental NLP components such as named entity recognition (NER) and disambiguation (NED)\textsuperscript{42,48,49}, entity linking\textsuperscript{51}, relation extraction\textsuperscript{47,48,52} and abbreviation recognition and disambiguation\textsuperscript{49}, and (2) components specific to clinical documentation (i.e. clinical insight extraction components) such as note section segmentation and classification\textsuperscript{53}, clinical semantic textual similarity (STS)\textsuperscript{50}, and assertion classification\textsuperscript{54}.

Clinical insight extraction components can be broadly categorized into (1) concept classification, (2) sentence classification, (3) sequential sentence classification, (4) semantic textual similarity, and (5) relation extraction. Concept classification components, such as assertion classification and medication change extraction, provide one or more categorial labels on a concept based on the surrounding context. Assertion classification uses a deep learning-based model that assigns each disease mention in a note with one of the following labels: present, absent, possible, hypothetical, conditional, associated with someone else; medication change extraction leverages multiple ML models that assign each medication mention in a note with labels corresponding to several dimensions: action (start, stop, increase, decrease), temporality (past, present, future), and certainty (certain, hypothetical, conditional). Sentence classification components assign one or more labels to a sentence in a clinical note; examples include binary classification models such as plan\textsuperscript{55}, follow-up, and goal extraction. Sequential sentence labeling components assign a label to each sentence in a given sequence, which are useful in capturing the inherent dependencies of sentences within a clinical note to assign the correct label for each sentence. For example, the note section classification model identifies the layout of a note with respect to its semantics by assigning one of a set of pre-defined section labels to each sentence (e.g. Review of Systems, Physical Exam). Clinical STS employs a combination of multi-task learning\textsuperscript{56} and fine-tuned pre-trained language models\textsuperscript{57,58} to compute the semantic equivalence between a pair of text snippets. Relation extraction components extracts entities from clinical notes and identifies relations between them. An example is ADE extraction which uses a joint deep learning model that extracts medications and related ADEs simultaneously.

Summarization components

Summarization components are high-level analytics built on top of the fundamental NLP and clinical insight extraction components. These include problem list generation, relation detection, disease-specific extractive summarization, consolidated medication timeline, and semantic find.
Evaluation of NLP components can be classified into intrinsic and extrinsic methods\(^6\). Intrinsic methods evaluate the functionality of a component against a predetermined gold standard, whereas extrinsic methods focus on the component’s contribution to the overall objective of an application and often requires a human-in-the-loop. As highlighted in by Pivovarov et al.\(^{13}\), extrinsic evaluation of NLP systems in a clinical setting is a challenging task due to reasons including (1) vendor EHR systems often do not support interaction with outside applications, and (2) hospitals often request evidence supporting the usefulness of an informatics system before investing resources for implementation. As a result of these constraints, here we report an evaluation of our system based on intrinsic methods, with an extrinsic evaluation of the overall application planned for the future.

As is common in information retrieval, recall (R), precision (P), and F\(_1\) score was used to evaluate our components. Recall, also known as sensitivity, is the number of true positives found given the total number of ground truth annotations. Precision, also known as positive predictive value, is the number of true positives given the total number of system predictions. F\(_1\) score provides a balanced measure of recall and precision. Specificity, also known as true negative rate, is not useful for our tasks as the number of true negatives are significantly higher than that of true positives, thereby yielding a less meaningful accuracy distinction.

**Problem list generation:** To extract an open-ended list of the patient’s active problems from the entire patient record, problem list generation was framed as a multi-label classification and ranking task. The model was based on an alternating decision tree, which used outputs from many upstream analytics (described above) as features, including but not limited to UMLS concepts identified by the NER/NED components and mapped to the CORE problem list subset of SNOMED CT, their assertions and distribution over the longitudinal record, and the degrees of alignment in temporal space between the identified candidate problems and their related medications, laboratory tests, and procedures. This method produced a human interpretable model, which is a desirable property in the clinical domain. To the best of our knowledge, this was the first successful open-ended problem list generation system\(^{15,44,45}\).

**Relation detection:** An ensemble approach with two supervised ML models was used to associate problems with related medications, laboratory tests, and procedures\(^9\). One method used features extracted from distributional semantics and UMLs; the second used features mined from historical, actual patient data. The results of relation detection were used both internally as features to support other components, and as a user interface functionality to sort structured data tables and highlight relevant data elements for a selected problem in the problem list.

**Disease-specific extractive summary:** Extractive summarization was modeled as a sequential sentence labeling problem, where each sentence in a given note is classified as being relevant (or not) to the management of the specified disease of interest. Outputs from upstream components, such as plan extraction and note section segmentation and classification, were leveraged in a linear-chain CRF model to automatically generate disease-specific summaries from clinical notes. As a post-processing step, boilerplate statements, which are often either not patient-specific or meant for purposes other than clinical care, are downweighted based on frequency analysis on all notes in our corpus. Next, deduplication of semantically equivalent sentences enabling the generation of more salient summaries was achieved by leveraging clinical STS. Details of the extraction summarization system was published in Liang et al.\(^{31}\).

**Consolidated medication timeline:** Data from structured medication orders, medications extracted from clinical notes and their related insights were aggregated to build a comprehensive medication timeline. In building this timeline, there were two key challenges: (1) some medications are only documented in unstructured data\(^5\) and (2) daily dosage, although present in many medication visualization designs\(^21,22\), is generally not available as a discrete field and requires understanding free text medication instructions (i.e. Sig). To address this gap, our system leveraged NER and NED components to identify medications and their attributes in clinical notes and Sigs, then normalized these attributes to calculate daily dosage\(^6\). To further enhance the timeline, outputs from clinical insight components (e.g. ADE and medication change extraction) were displayed as popovers on the timeline for the relevant medication.

**Semantic find:** Semantic find is an NLP-based search that leverages UMLS and distributional semantics to perform different kinds of search, such as conceptual search, which leverages UMLS to match clinically equivalent concepts (e.g. lisinopril/Zestril) and ISA relations to match more specific (e.g. cancer/sarcoma) or more general (e.g. gi bleed/hemorrhage) variations of the search term, and associative search, which uses Latent Semantic Analysis\(^51\) to measure the probability of an association between the search term and target term, and returns concepts with a score above 0.5 (e.g. asthma/wheezing). Details for semantic find are available in Prager et al.\(^32\).

**Evaluation metrics**
Results

Intrinsic evaluation

Table 1 presents the precision, recall, and $F_1$ score for the summarization components, namely problem list generation, relation detection, disease-specific extractive summarization, consolidated medication timeline (reported separately for ADE and medication change extraction), and semantic find.

Table 1. Precision, recall and $F_1$ score for problem list generation, relation detection, disease-specific extractive summarization, adverse drug event extraction, medication change extraction, and semantic find.

<table>
<thead>
<tr>
<th>Component</th>
<th>Type</th>
<th>Precision</th>
<th>Recall</th>
<th>$F_1$ score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Problem list generation</td>
<td>(all problems)</td>
<td>0.84</td>
<td>0.66</td>
<td>0.74</td>
</tr>
<tr>
<td>Relation detection</td>
<td>Problem – Medication</td>
<td>0.82</td>
<td>0.75</td>
<td>0.78</td>
</tr>
<tr>
<td>Relation detection</td>
<td>Problem – Procedure</td>
<td>0.76</td>
<td>0.69</td>
<td>0.71</td>
</tr>
<tr>
<td>Relation detection</td>
<td>Problem – Laboratory Test</td>
<td>0.85</td>
<td>0.74</td>
<td>0.79</td>
</tr>
<tr>
<td>Disease-specific extractive summarization</td>
<td>Hypertension</td>
<td>0.74</td>
<td>0.58</td>
<td>0.65</td>
</tr>
<tr>
<td>Disease-specific extractive summarization</td>
<td>Diabetes mellitus</td>
<td>0.74</td>
<td>0.62</td>
<td>0.67</td>
</tr>
<tr>
<td>Disease-specific extractive summarization</td>
<td>Hyperlipidemia</td>
<td>0.67</td>
<td>0.47</td>
<td>0.55</td>
</tr>
<tr>
<td>Disease-specific extractive summarization</td>
<td>Heart failure</td>
<td>0.74</td>
<td>0.50</td>
<td>0.60</td>
</tr>
<tr>
<td>Disease-specific extractive summarization</td>
<td>COPD</td>
<td>0.69</td>
<td>0.53</td>
<td>0.60</td>
</tr>
<tr>
<td>Adverse drug event extraction</td>
<td>Medication – Adverse Event</td>
<td>0.61</td>
<td>0.56</td>
<td>0.58</td>
</tr>
<tr>
<td>Medication change event extraction</td>
<td>Action</td>
<td>0.81</td>
<td>0.77</td>
<td>0.79</td>
</tr>
<tr>
<td>Medication change event extraction</td>
<td>Temporality</td>
<td>0.79</td>
<td>0.79</td>
<td>0.79</td>
</tr>
<tr>
<td>Medication change event extraction</td>
<td>Certainty</td>
<td>0.78</td>
<td>0.76</td>
<td>0.77</td>
</tr>
<tr>
<td>Semantic find</td>
<td>(all search terms)</td>
<td>0.87</td>
<td>0.87</td>
<td>0.87</td>
</tr>
</tbody>
</table>

ProSPER application

ProSPER consists of (1) a problem-oriented patient summary centered around an auto-generated problem list, and (2) disease-specific views for several chronic conditions. The problem-oriented patient summary allows users to gain a quick overview of the patient, while the disease-specific views allow a deeper dive into a specific problem of interest.

The problem-oriented patient summary presents (1) an auto-generated problem list, (2) aggregated clinical data by type, (3) relations between problems and other data aggregates, and (4) a semantic search over the entire EHR. Figure 2 shows the user interface which presents several panels, one for each data type. When a problem is selected, related medications, laboratory results, and procedures are highlighted and brought to the top for easier review.

Figure 2. Problem-oriented patient summary with diabetes mellitus selected.
The disease-specific view allows further investigation of a problem by presenting all information relevant to the disease of interest on one screen. In this view, encounters in which the specified disease was discussed are displayed on a timeline, and an extractive summary of each of these encounters is produced containing only the most important information relevant for managing the specified disease. Relevant medications, test results and their trends are also displayed in timeline form, allowing users to identify potentially informative encounters, and also infer relationships between various elements of the patient record (e.g. medication prescribed after an increase in a particular laboratory result). Figure 3 shows the disease-specific view for diabetes mellitus in a patient.

Figure 3. Disease-specific view for diabetes mellitus.

Disease-specific extractive summaries for each encounter are displayed by hovering over icons in the encounter timeline, with the summary of the most recent encounter in the right panel for easy viewing. Orange indicators on the medication timeline represent medication insights extracted from clinical notes and are displayed as popovers. Semantic find can be launched using the search box on the top right if the user needs additional information.

Discussion

Challenges in developing an NLP-based application for use in a real-world clinical setting

There is a significant gap between what low-level NLP components can do as compared to the mental work physicians perform during chart review. Previous studies on applying NLP in the clinical domain either focused on advancing the state-of-the-art on a single task or using existing NLP toolkits with more emphasis placed on data visualization. This paper tries to bridge the gap between what low-level NLP components can provide versus what physicians want, by leveraging ML and NLP in all aspects of a chart review scenario.

Several components within our system have been previously published to advance the state-of-the-art on publicly available datasets. However, for many components there are no publicly available datasets, necessitating use of in-house datasets and ground truth. An additional challenge is the need for both structured and unstructured data in creating ground truth for many high-level tasks such as summarization.

Although NLP applications have reported good accuracy, studies have shown that these applications still have low adoption rates in clinical practice, largely due to uncertainty about minimal performance requirements rather than performance-related issues. The interpretation of a specific component’s accuracy in terms of the overall usefulness of a system as measured by an extrinsic evaluation still warrants more research attention.

How does ProSPER address physician challenges in reviewing EHRs?

To understand how ProSPER can support physicians in their review of EHR data, we discuss how specific components within the system address the five information hazards identified by Beasley et al.: information overload, information underload, information scatter, information conflict, and erroneous information. Information overload occurs when

769
there is too much data for the user to review, organize, synthesize, and act on. One way ProSPER mitigates this problem is by disease-specific extractive summarization, which leverages components such as note section classification and clinical STS to remove uninformative or redundant text\textsuperscript{24,64}, to produce a succinct summary of a clinical note. Information underload describes the lack of necessary information, such as missing information due to incomplete patient records, or information that is so difficult to find and for all intents and purposes, “missing”. ProSPER targets information underload in several ways, including semantic find, which searches the entire EHR based on a user-provided query, and ADE extraction, which surfaces rare yet important clinical events that have great impact on patient safety. Information scatter describes the problem of having information located in multiple places, thereby requiring additional effort to reconcile this information. The idea of a problem list was conceptualized as a way to address this issue by capturing all of the patient’s important health problems in one centralized location, but in practice it is rarely well-maintained rendering it mostly unusable\textsuperscript{65}. ProSPER supports problem list reconciliation through its problem list generation component, which has been validated by physicians in a small pilot study\textsuperscript{66}. The last two information hazards, information conflict and erroneous information, can be discussed together as conflicts within EHR documentation can be partially attributed to erroneous information. In providing a consolidated medication timeline containing information from both structured and unstructured sources, ProSPER allows users to more easily identify and resolve discrepancies in medication data between structured and unstructured data.

**Limitations and future work**

We acknowledge several limitations to this work. First, the results presented in this paper are based on an intrinsic evaluation of the NLP components using precision, recall, and F\textsubscript{1} score to compare against ground truth created by physicians or medical students. However, these metrics do not fully capture the nuances of what should or should not be included in a problem-oriented patient summary. For instance, a single missed ADE by the system may not have a significant impact when measured using the proposed quantitative metrics, but it may be critically important to patient management and should never be missed. Second, these results are specific to the datasets, which contain only EHRs from two healthcare organizations. Ground truth generation and evaluation on a more diverse dataset is needed to better understand the effectiveness of these approaches. Third, while some of the components were previously evaluated and published using publicly available datasets, others were developed using proprietary datasets not available to the research community. In addition to continuing to adapt and innovate novel methodologies to improve individual NLP components, which in-turn can benefit the overall application, we plan to take the following steps to address the limitations above:

1. Use extrinsic measures to capture the usefulness of ProSPER for practicing physicians at the point-of-care.
2. Create ground truth annotations on publicly available datasets such as MIMIC-III\textsuperscript{66} and organize challenges to invite community-driven efforts.

**Conclusion**

We present ProSPER, a novel problem-oriented summary of the patient record that leverages 3 levels of NLP informatics, including performing well-studied fundamental NLP tasks such as NER, extracting clinically meaningful events from clinical notes such as ADE, generating an active problem list from the heterogeneous, longitudinal patient record, and creating disease-specific extractive summaries from clinical notes. Components underlying ProSPER target specific information hazards within the EHR that lead to excess physician workload and mental exhaustion. Our work demonstrates the need to leverage holistic information in EHRs to build a comprehensive summarization application, and the potential for NLP-based applications to support physician and improve overall clinical care.

**References**

Extraction of Active Medications and Adherence Using Natural Language Processing for Glaucoma Patients

Wei-Chun Lin, MD, MS¹, Jimmy S. Chen, MD², Joel Kaluzny, MD³, Aiyin Chen, MD³, Michael F. Chiang, MD⁴, Michelle R. Hribar, PhD¹
¹Medical Informatics & Clinical Epidemiology, ²School of Medicine, ³Ophthalmology Oregon Health & Science University, Portland, OR, ⁴National Eye Institute, Bethesda, MD

Abstract

Accuracy of medication data in electronic health records (EHRs) is crucial for patient care and research, but many studies have shown that medication lists frequently contain errors. In contrast, physicians often pay more attention to the clinical notes and record medication information in them. The medication information in notes may be used for medication reconciliation to improve the medication lists' accuracy. However, accurately extracting patient's current medications from free-text narratives is challenging. In this study, we first explored the discrepancies between medication documentation in medication lists and progress notes for glaucoma patients by manually reviewing patients' charts. Next, we developed and validated a named entity recognition model to identify current medication and adherence from progress notes. Lastly, a prototype tool for medication reconciliation using the developed model was demonstrated. In the future, the model has the potential to be incorporated into the EHR system to help with real-time medication reconciliation.

Introduction

The rapid adoption of electronic health records (EHRs) has generated large-scale clinical data that has been re-used for many purposes, including patient phenotyping, pharmacovigilance, comparative effectiveness research, clinical decision support, and quality improvement and research. Although secondary use of EHR shows many benefits such as improved healthcare quality, reduced healthcare costs, and effective clinical research, there are many challenges that still need to be addressed. One of the biggest challenges is the accuracy and completeness of EHR data, specifically medication information.

The accuracy of medication data is crucial for patient safety, quality of care, and clinical research. Inaccurate or incomplete medication records can lead to polypharmacy, adverse medication interactions, and decreased data reliability in research. The medication list is a structured record of a patient's medication data which is populated automatically by electronically prescribed medications or manually through medication reconciliation. However, the EHR system may not always capture medication data correctly or prevent errors in the medication list. Previous studies have shown that medication lists frequently contain errors, including duplicated documentation of medications, outdated discontinued prescriptions in the medication list, and missing medications prescribed elsewhere. In addition, prior studies show that physicians direct very little attention to EHR medication lists, and instead spend most time reviewing the impression and plan section. It seems reasonable to expect that medications recorded in narrative notes are more reliable and can be helpful with medication reconciliation. Medication reconciliation is a process to create and maintain patients' most current and accurate list of medications.

However, manual reviewing progress notes for medication data extraction in EHR is time-consuming and labor-intensive. Natural language processing (NLP) is a promising strategy for capturing medication information from the free-text progress note. With advancements in machine learning and the large text corpora available in EHR, NLP has been successfully used to process free-text EHR data, for deep contextualized word representations, information extraction, and semantic analysis. Named entity recognition (NER) is a sub-task of information extraction, which seeks to identify words or phrases into pre-defined categories with specific labels. Over the past years, the NER technique has been applied to extract medication information, such as drug names, frequency, dosage, adverse drug events, adherence, etc., from free-text documents. For example, a conditional random field (CRF) model was used to develop a NER model to detect medication attributes and adverse drug events. Also, bidirectional long short-term memory (LSTM) model was used for named entity recognizing for medication information. More recently, pre-trained deep learning models were widely used for biomedical information extraction. However, to our knowledge, there has not been a well-developed NLP tool to identify a list of current medications for a specific disease such as glaucoma and help with medication reconciliation.
The purpose of this study is to develop a NER model for extracting patients’ current ophthalmologic medication and adherence from free-text notes for glaucoma patients. Glaucoma is characterized by progressive degeneration of the optic nerve and irreversible visual field loss, and it is the leading cause of irreversible blindness worldwide.\textsuperscript{33} The majority of glaucoma patients are treated using medical therapy, and the accuracy of medication documentation is crucial in glaucoma management.\textsuperscript{34} However, the accuracy of glaucoma medication documentation is unclear. In addition, glaucoma patients’ medication non-adherence rate has been reported to vary from 24% to 59%.\textsuperscript{35, 36} Therefore, a reliable method to assess glaucoma patients’ current ophthalmologic medication and adherence is needed.\textsuperscript{37} Finally, the reliability of medication data is important for glaucoma research, such as prediction models for disease progression. In this study, we first manually reviewed patient charts for discrepancies in medication documentation between medication lists and progress notes. Next, we trained and tested a NER model for extracting current medication from progress notes and evaluated its accuracy. Finally, we demonstrated an approach for medication reconciliation using the NER model on small sample progress notes.

Methods

This study was approved by the Institutional Review Board at Oregon Health and Science University (OHSU). OHSU is a large academic medical center in Portland, Oregon. This study was conducted at Casey Eye Institute, OHSU’s ophthalmology department serving all major ophthalmology subspecialties. The department performs over 130,000 outpatient examinations annually and is a major referral center in the Pacific Northwest and nationally. In 2006, OHSU implemented an institution-wide EHR (EpicCare; Epic Systems, Verona, WI) to handle all ambulatory practice management, clinical documentation, order entry, medication prescribing, and billing.

The study contains three phases (1) Explore medication discrepancies between the medication list and the progress note for glaucoma by manually reviewing charts; (2) Develop a NER model to extract patients’ current ophthalmologic medication and medication adherence from progress notes for glaucoma patients and (3) Apply the NER model to perform medication reconciliation.


Progress notes and medication list data from EHR were extracted for 150 randomly selected Casey Eye Institute patients with encounter ICD10 diagnosis codes related to glaucoma from January 23, 2019, to September 28, 2020. The patient’s most recent office visit notes were manually reviewed by three independent reviewers. The medications recorded in the narrative notes were abstracted and compared to the EHR medication list at the time of visit. All ophthalmologic medications and over-the-counter (OTC) medications (e.g., artificial tears) were collected. All medications listed in the notes but not on the medication list or vice versa were labeled. Cross-validation among the three reviewers was conducted by using a subset of 20 encounter notes (96.4% agreement).

2. NER Model for Extracting All Ophthalmic Medications

We sampled a dataset with 507 progress notes from office visits at the Casey Eye Institute from January 01, 2019, to December 31, 2019, with encounter ICD10 codes associated with glaucoma. The dataset was constructed by random stratified sampling from all ophthalmology visits according to the department and primary provider name. The documents were manually annotated for nine categories: Drug Name, Route, Frequency, Dosage, Strength, Duration, Adverse Drug Event (ADE), Adherence, and Current Medication Use. All medication names, including generic names, brand names, and abbreviations, were sourced from publicly available online resources and glaucoma specialists. An open-source tool (Doccano; Open source: Doccano; 2018) was used to annotate the documents.\textsuperscript{38} Due to the limited number of ADE entities, we discarded this category and kept the other eight entities. Figure 1 displays an example of the annotation. The annotated dataset was randomly split into 75% for training and 25% for testing. A 10% randomly sampled subset of documents from the training data was used as a validation set for turning the hyperparameters. Table 1 presents the description of the datasets and annotation statistics.
Figure 1. Example of note annotation by an open-source tool. Medication drug name, strength, dosage, frequency, route, duration, current use, and adherence are identified.

We used named entity recognition, a sub-type information extraction technique, to extract medication information and adherence from clinical notes. The NER model was developed in Python 3.7.6 using the spaCy library. The spaCy library is a free open-source library for NLP. The architecture of spaCy’s NER model is based on convolutional neural networks which uses a word embedding strategy using sub-word features and "Bloom" embeddings. In this study, the training task contains 200 epochs with experiments with multiple hyperparameter settings. Different learning rates (initial at 1e−2, 1e−3, 2e−3, 1e−4, 2e−4) were tested and adjusted by two optimizers: Adaptive Moment Estimation (Adam) and stochastic gradient descent. We use a decaying dropout rate (0.5 - 0.35; 1e−3) to avoid overfitting. Also, we experimented with different batch compounding sizes and regularization schemes to optimize the model. The results of the NER model’s extraction for the test set were determined by comparing the manually annotated and the NER model’s extracted entities. The model performance was evaluated by using F1 score, precision, recall, and the micro-averaged score, which aggregates the contributions of all categories to calculate the average metrics.

Table 1. Distribution of annotated entities and number of progress notes in training and testing datasets.

<table>
<thead>
<tr>
<th>Named Entities</th>
<th>Train</th>
<th>Test</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drug</td>
<td>2029</td>
<td>505</td>
<td>2534</td>
</tr>
<tr>
<td>Frequency</td>
<td>1722</td>
<td>411</td>
<td>2133</td>
</tr>
<tr>
<td>Route</td>
<td>1666</td>
<td>371</td>
<td>2037</td>
</tr>
<tr>
<td>Dosage</td>
<td>201</td>
<td>40</td>
<td>241</td>
</tr>
<tr>
<td>Duration</td>
<td>35</td>
<td>15</td>
<td>50</td>
</tr>
<tr>
<td>Strength</td>
<td>168</td>
<td>31</td>
<td>199</td>
</tr>
<tr>
<td>Adherence</td>
<td>132</td>
<td>48</td>
<td>180</td>
</tr>
<tr>
<td>Current Medication Use</td>
<td>725</td>
<td>185</td>
<td>910</td>
</tr>
<tr>
<td>Number of Notes</td>
<td>381</td>
<td>126</td>
<td>507</td>
</tr>
</tbody>
</table>
3. Medication Reconciliation Using NER Model for Current Medications

Finally, we developed a prototype medication reconciliation tool using the optimized NER model. For this purpose, we are focusing only on medications the patient is currently using as documented in progress notes. Figure 2 demonstrates an example of medication reconciliation using our prototype tool. First, our NER model extracted the patient’s medications and "Drug Use" label from the 150 sample progress notes which were manually reviewed in phase 1. The “Drug Use” label identified which medications that the patient was currently taking. Next, the current medications were standardized based on RxNorm Ingredient (IN). Finally, the standardized medications were compared to the manually identified medications from phase 1. Both ophthalmologic medications and over-the-counter (OTC) medications (e.g., artificial tears) were included. All medications listed in the notes but not on the medication list or vice versa were flagged.

![Figure 2](image_url)

**Figure 2.** Example of medication reconciliation using the developed NER model

**Results**


The randomly sampled 150 patients’ notes and medication lists contained a total of 450 medications, including glaucoma eye drops, mydriasis eye drops, antimicrobials, corticosteroids, and OTC medications. Prescription medications were most common \((n = 355; 79\%)\), followed by OTC medications \((n = 95; 22\%)\). Around 57% of patients had at least one medication mismatch for all categories in their records. However, only 36% of patients had at least one medication mismatch for prescription medications \((n = 298)\). Nearly 66% of medications could be reconciled between the progress notes and medication list. Around 34% \((n = 152)\) of medications are mismatched for various reasons, including medications prescribed by clinicians from different institutions, medications with duplicated prescriptions, medications that were prescribed and entered in the medication list but not recorded in the progress note, and old medications that were not discontinued in the medication list. Figure 4 displays the distribution of medication mismatches among the two categories in the EHR by location. The most frequent mismatch was found with prescription medications \((55\%)\) followed by the OTC medications \((45\%)\). The OTC medications were more commonly recorded in the progress notes but not entered into the medication list. In contrast, mismatched prescription medication more often appeared in the medication list but not in the progress notes.
Figure 3. Medication documentation mismatches were stratified based on the number of mismatches that occurred per patient for prescription (blue), OTC (green), and all medications (yellow).

Figure 4. Summary of medication mismatches across 150 patients.
2. NER Model for Extracting Current Ophthalmic Medication
The custom NER model was trained with 381 progress note documents that were manually annotated with eight named entities and then tested on 126 progress notes. Table 2 presents the overall micro-averaged and per-entity performance for the optimal NER model on test data (126 progress notes). The overall performance of the NER model across all categories was F1 score = 0.955, Precision = 0.951, and Recall = 0.957. Higher performance was observed on medication-related entities: Drug, Name, Route, Frequency, Dosage, and Strength, compared to patient’s behavior-related entities: Adherence and Current Medication Use. An error analysis was performed for false negative and positive on Drug Name, Adherence, and Current Medication Use to recognize the source of error predictions. Several causes of errors were identified, such as different wordings for medication adherence, mislabeled current medication use and drug name due to similar sentence structure, eye exams or warm compress mislabeled as drug name, and misclassification when entity information was contained in a short sentence. (Table 3).

Table 2. The results of the NER model on the test dataset

<table>
<thead>
<tr>
<th>Entities</th>
<th>Performance on Test Data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
</tr>
<tr>
<td>Drug</td>
<td>0.971</td>
</tr>
<tr>
<td>Frequency</td>
<td>0.972</td>
</tr>
<tr>
<td>Route</td>
<td>0.948</td>
</tr>
<tr>
<td>Dosage</td>
<td>0.987</td>
</tr>
<tr>
<td>Duration</td>
<td>1.000</td>
</tr>
<tr>
<td>Strength</td>
<td>0.969</td>
</tr>
<tr>
<td>Adherence</td>
<td>0.803</td>
</tr>
<tr>
<td>Current Medication Use</td>
<td>0.899</td>
</tr>
<tr>
<td>Average (micro)</td>
<td>0.951</td>
</tr>
</tbody>
</table>

Table 3. Error analysis from NER predictions related to Drug Name, Current Medication Use, and Adherence labels

<table>
<thead>
<tr>
<th>Error category</th>
<th>Example</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mislabeled Current Medication Use</td>
<td>“Urgent add on - Last seen Dr. X on X/X/XXXX”</td>
<td>Unexplained error, “Urgent add on” was labeled as Current Medication Use</td>
</tr>
<tr>
<td></td>
<td>“encouraged PFATs at least BID OU - discussed to space at least 5 mins from glaucoma drops”</td>
<td>“Encouraged” was mislabeled as Current Medication Use due to the similar sentence structure</td>
</tr>
<tr>
<td>Mislabeled Adherence</td>
<td>“- History inconsistent drop adherence”</td>
<td>There are many different wordings for medication adherence, and Adherence label was not assigned</td>
</tr>
<tr>
<td></td>
<td>“No eye pain/discomfort but patient admits to forgetting his drops frequently.”</td>
<td></td>
</tr>
<tr>
<td>Mislabeled Drug Name</td>
<td>“Cont warm compresses BID ou”</td>
<td>“warm compresses” was mislabeled as Drug Name due to a similar sentence structure</td>
</tr>
<tr>
<td></td>
<td>“Vision has been good. Just using OTC readers.”</td>
<td>“Using” was mislabeled as Current Medication Use, and OTC readers was mislabeled as Drug Name due to the similar sentence structure</td>
</tr>
</tbody>
</table>
3. Medication Reconciliation Using the NER Model

The prototype medication reconciliation tool identified 408 current medications from the 150 progress notes that were manually reviewed in phase 1. After standardizing the medications to RxNorm, 14 medications were removed for a final list of 394 medications. Among the 394 medications, there were 379 medications matched with the manually abstracted current medications. The prototype tool achieved a good performance of F1 score = 0.969, Precision = 0.959, and Recall = 0.979.

Discussion

In this study, we explored medication discrepancies in the EHR data and evaluated the performance of a custom NER model's applicability to extract current medication for glaucoma patients. We also used the developed NER model in a proof-of-concept application to perform medication reconciliation in a subset of our patients. The key findings from our study were (1) Medication discrepancies in patient charts were found to be present in a large proportion of office visits; (2) The custom NER model can accurately extract current medication and adherence for glaucoma patients; (3) The NER model can be used to reconcile the medication documentation.

The first key finding is that medication discrepancies were found to be present in a large proportion of office visits. Our study shows that approximately twenty percent of medications prescribed to glaucoma patients had at least one discrepancy between the medication list and the progress note. Overall, more than one-third of patients in this study had at least one medication mismatch between both data sources. These inconsistencies in the EHR medication records may increase the risk of medication errors and affect the reliability of research that relies on this data. These findings are similar to other studies, including a study for microbial keratitis demonstrating 76.9% of medication agreement between progress notes and medication lists and another study for inflammatory bowel disease reporting 78.6% of medications agreement between clinical narrative and medication list. The findings from these studies indicate that the accuracy of the medication list is a common problem. An accurate tool for medication reconciliation of medication lists and further qualitative studies to understand the causes of medication data discrepancies is needed.

The second key finding is that our NER model can accurately identify current medication and adherence from progress notes from outpatient glaucoma visits. In our study, the model reached a micro-averaged F1 score of 0.955 across all categories. The NER model was developed to recognize eight categories from free-text progress notes, including drug names (including generic, brand, and abbreviation names), the route of administration, prescription frequency, the dosage of the drugs, drug strength, duration, medication adherence, and current medication use. The NER model could accurately identify medication-related entities (except duration) but showed lower performance on patient behavior-related entities, such as adherence and current medication use. The difference could be ascribed to the limited number of training cases and the higher variety of wordings. As shown in Table 1, there are only 35 annotated duration entities and 132 annotated adherence entities in the training data. In addition, the words and phrases to indicate adherence and current medication use are various, and some of these phrases are located in different sentences than the medications. Nevertheless, the most common error of drug name identification is mislabeling other terms such as “warm compress” or “OTC readers” as a drug name due to similar sentence structure. For example, “warm compress left eye PRN” or “Vision has been good. Just using OTC readers.” In these cases, these mislabeled drugs will easily be filtered out of the results in practice during the conversion to RxNorm names.

Finally, our NER model can be used to reconcile medication documentation. As shown in the phase one study, we can manually abstract the medication records from their progress notes to compare with their medication list. Similarly, the NER model was able to recognize common medications as well as identify text related to current medication use. This is the first study, to our knowledge, to develop NLP models to recognize current medication use from free-text progress notes. With the ability to identify the current medication use, we are able to capture the whole picture of current medications for the target patients and reconcile it with their medication list. As previously mentioned, the medication reconciliation between progress notes and medication lists was only reported from 76.9% to 79.6% for three different diseases, including microbial keratitis, inflammatory bowel disease, and glaucoma. And more than one-third of patients had at least one discrepancy for ophthalmic prescription medications. In our study, the NLP tool can correctly identify current medications for glaucoma patients on 150 sample progress notes (F1 score = 0.969). Figure 2 displays an example of medication reconciliation using the NLP tool. In this prototype tool, we focused on reconciling the drug names since physicians did not always record the other attributes, such as route, frequency, and dosage along with the medications. In future work, we plan to extend this medication reconciliation method to use the
information from both narrative progress notes and medication lists to construct a current medication list for glaucoma patients.

Our study has limitations future work may address. First, some of the entities are naturally less frequently recorded in the progress notes that affect the performance of the NER model. For example, text related to drug duration appeared much less frequently than other entities, such as drug name, route, and frequency. Thus, it is challenging to train the model correctly recognize these entities. A similar finding was reported in another study.31 Second, the model was trained on a set of notes for glaucoma patients from a single institution; it is unclear if the model can be generalizable to other subspecialties within ophthalmology or other healthcare systems. Finally, the application of the custom NER model for medication reconciliation is a proof of concept. We conducted the test of medication reconciliation using the NER model on a limited number of samples. Our intention is to extend and replicate these study methods to different specialties and institutions to increase the generalizability of our model. In the future, the custom model could be incorporated into the EHR system to help with medication reconciliation.

**Conclusion**

Discrepancies in medication documented in the medication and in progress notes were observed in more than one-third of encounters for glaucoma patients. Inaccurate medication lists in the EHR may affect the reliability of the research or clinical decision support using this data. Since physicians often record current medication information in the progress notes this data could be used for medication reconciliation. In this study, we developed an NLP model to accurately identify current medication information from free-text EHR data that can be applied to perform automated medication reconciliation; the performance of the model is similar to the best performing published NLP models for medication extraction studies.25-31,46 This has implications in improving the data quality and usefulness for medication data in both research and clinical care.

**Acknowledgments**

Supported by grants T15LM007088, 1R21EY031443-01, and P30EY0105072 from the National Institutes of Health (Bethesda, MD) and unrestricted departmental support from Research to Prevent Blindness (New York, NY). MFC was a consultant for Novartis (Basel, Switzerland) and previously an equity owner in InTeleretina LLC (Honolulu, HI), and received research support from Genentech (San Francisco, CA) and the National Science Foundation (Alexandria, VA).

**References**


Development and Evaluation of an Automated Approach to Detect Weight Abnormalities in Pediatric Weight Charts

Lei Liu, BS1,2, Danny T.Y. Wu, PhD, MSI1,3, S. Andrew Spooner, MD2,3, Yizhao Ni, PhD2,3

1Department of Biomedical Informatics, College of Medicine, University of Cincinnati, Cincinnati, OH; 2Division of Biomedical Informatics, Cincinnati Children’s Hospital Medical Center, Cincinnati, OH; 3Department of Pediatrics, College of Medicine, University of Cincinnati, Cincinnati, OH

Abstract

Inaccurate body weight measures can cause critical safety events in clinical settings as well as hindering utilization of clinical data for retrospective research. This study focused on developing a machine learning-based automated weight abnormality detector (AWAD) to analyze growth dynamics in pediatric weight charts and detect abnormal weight values. In two reference-standard based evaluation of real-world clinical data, the machine learning models showed good capacity for detecting weight abnormalities and they significantly outperformed the methods proposed in literature (p-value<0.05). A deep learning model with bi-directional long short-term memory networks achieved the best predictive performance, with AUCs ≥0.989 across the two datasets. The positive predictive value and sensitivity achieved by the system suggested more than 98% screening effort reduction potential in weight abnormality detection. Consequently, we hypothesize that the AWAD, when fully deployed, holds great potential to facilitate clinical research and healthcare delivery that rely on accurate and reliable weight measures.

Introduction

Body weight, as an important parameter in pediatrics, is widely utilized to track growth and development of children and to calculate drug doses in clinical settings1–4. It is also an important variable for secondary data analysis in clinical research such as computerized phenotyping5. In current practice, body weight is usually measured and entered into the electronic health records (EHRs) by clinical staff manually, and can be recorded incorrectly in various ways, including improper operation of weight scale, typing errors, conversion errors (e.g., weight scale in pounds, and documentation in kilograms), and weight estimations instead of actual measurements4. Inaccurate body weight data have a high risk of causing patient harm; literature studies have reported that 18-22% of medication errors are resulted from “improper dose/quantity” in pediatrics, which is significantly higher than that in adult settings2,3. The erroneous data also hinder the full utilization of EHRs for research purpose due to error propagation in downstream analysis.

Manual chart review is currently standard practice in correcting weight abnormalities, which is labor-intensive and impractical to perform in busy clinical settings. In particular, capturing weight abnormalities with high accuracy can be difficult, especially for neonates, and growing children who have acute or chronic medical conditions. There is a critical need to develop an accurate and cost-effective approach for detecting abnormalities from pediatric weight charts to prevent weight-based dosing/medication errors. The approach can also be applied to cleaning abnormal weight values from EHRs to provide high-quality data for subsequent research.

By using recent informatics technologies, several approaches have been proposed to identify abnormal weights from patient charts4. One established method is developed by the Centers for Disease Control and Prevention (CDC), which standardizes weight points with z-score normalization and identifies outliers based on their standard deviations from the mean7. Another computerized approach is proposed by Children’s Hospital of Philadelphia, which compares standard deviation of a weight point against a weighted moving average in the chart to identify abnormal values8. In our earlier study, we also develop a regression approach to model weight trend in a chart and determine if a weight point is an outlier9. However, all the methods are rule-based, where empirical thresholds are developed based on subjects’ age and gender to identify abnormal values. Even though they are easy to implement, the approaches have low detection capacity, particularly for complex weight charts.

Machine learning is a field of artificial intelligence that utilizes computerized algorithms to learn the relation between, and make prediction on, sets of data. The technologies have been widely used on a variety of clinical decision support tasks, including patient clinical status detection, workflow optimization, and computerized phenotype discovery10-12. More recently deep learning, a branch of machine learning focusing on developing neural network-based algorithms, has gained increasing popularity in clinical informatics and has been applied to analyze sequence data such as clinical narratives13-15. Nevertheless, no studies have used machine learning for analyzing patient weight charts.
Our research is specifically directed at developing an accurate and scalable informatics-based solution, an automated weight abnormality detector (AWAD), to identify errors in pediatric weight charts. In our earlier studies, we developed a visual annotation tool to enable large-scale annotation of weight abnormalities. To take the next step, this study focused on developing an automated approach to analyze patient charts and identify abnormal weight values. We hypothesized that with using state-of-the-art machine learning and deep learning technologies, the AWAD could detect weight abnormalities for individual charts with high sensitivity and specificity. The study is the first, known to us, to investigate detection of weight abnormalities in a large-scale via machine learning technologies.

**Materials and methods**

Figure 1 diagrams the overall processes of the study. We first collected and selected weight charts from the institutional EHRs (process 1 in Figure 1). Annotation was then performed to identify abnormal weight values for individual charts (process 2). Features were extracted from each weight chart to capture weight characteristics and growth dynamics (process 3), which were then fed into machine learning- and deep learning-based algorithms to detect abnormal weight values (process 4). Finally, a separate annotated weight set was applied to assess the generalizability of the developed algorithms (process 5).

**Data collection and weight chart selection**

A total of 4.3 million weight points was collected from the EHRs for all 347,056 patients visiting Cincinnati Children’s Hospital Medical Center (CCHMC) between 2010 and 2018. Use of the de-identified dataset was approved by the University of Cincinnati Intititional Review Board (study ID: 2017-2075). Following our earlier study, we excluded three types of weight points from analysis: 1) weight points documented in the first 24 months (1,137,410 points from 31,876 patients) because newborns could have more complex weight changing patterns, 2) weight points recorded after 240 months (17,211 points from 7,932 patients), and 3) all weight points from patients with less than four weight measures (213,388 points from 87,013 patients) to ensure sufficient information in each weight chart. After data exclusion, a secondary polynomial regression model was built to identify charts with potentially abnormal weight values based on five parameters including 1) maximum absolute residuals, 2) median absolute residuals, 3) root mean square errors, 4) maximum ratio of absolute residuals to fitted values, and 5) mean ratio of absolute residuals to fitted values. The regression model identified 107,336 candidate charts with at least one weight point outside its 99% confidence intervals. A set of 15,000 charts (denoted by MAIN dataset) were randomly sampled from the 107,336 dataset for weight abnormality annotation.

**Weight abnormality annotation**

We recruited domain experts with appropriate medical training and knowledge to review the selected weight charts and identify abnormal weight values. The annotation guideline suggested 1) weight points should be evaluated retrospectively based on all available points in a chart, and 2) weight points should be labeled as abnormal based on their clinical importance and risk for leading to patient harm. By using the guideline and a visual annotation tool developed in-house, the 15,000 weight charts were annotated by 18 domain experts to create a reference-standard set of weight abnormalities. During annotation, each expert was assigned to a random sample of 2,500 patient charts such that each chart was reviewed by three annotators. An example chart with annotation is presented in Figure 2. All weight points were considered normal and defaulted to blue. Abnormal points with high clinical importance were labeled as red, while potentially abnormal values requiring second opinions were labeled as orange.

A scoring approach was applied to summarize the annotations, with red scoring two, orange one and blue zero. A weight point was considered abnormal if sum of scores from the three annotators were greater than two (i.e., at least one red and one orange decisions or three orange decisions). Weight points scoring two were further reconciled by
our clinical champion (Dr. Spooner) to determine their abnormality. Adjacent data points with identical age and weight values were reviewed manually to avoid diluting the effects of predictors by contradictory outcome labels. The inter-rater agreement (IRA) was calculated using Fleiss’ Kappa.

**Figure 2.** An example chart with weight abnormality annotation. The weight characteristics presented in the box were available to annotators during annotation. Color codes: 1) blue represents normal weight entries, 2) red represents abnormal weight points with high clinical importance, and 3) orange represents potentially abnormal points requiring second opinions.

**Weight feature extraction**

For each weight point, nine variables were generated to capture weight characteristics and growth dynamics: 1) subject weight in kilograms, 2) subject age in years, 3) subject sex, 4) the Box-Cox transformation, median and generalized coefficient of variation- (LMS-) based z-score according to subject sex and age, 5) modified LMS-based z-score using the weight-for-age data from a reference population provided by CDC to identify extreme weight values, 6) percentage of the population that was below a weight value (denoted by percentiles), 7) absolute age difference from the immediate previous weight point, 8) absolute weight difference from the immediate previous weight point, and 9) absolute z-score difference from the immediate previous weight point. The numerical variables were standardized with z-score normalization. The categorical variable (i.e., sex) was binarized with a dummy variable to avoid linear dependencies. After feature extraction each weight point was represented by a nine-dimensional numerical vector.

**Traditional machine learning classifiers**

We model detection of weight abnormalities as a binary-class classification and implemented five machine learning classifiers including LR with L1 and L2 normalization, support vector machines with linear (SVM-L), polynomial (SVM-P) and radial basis function (SVM-R) kernels, decision trees (DTs), random forests (RFs), and one-layer artificial neural networks (aNNs).

**Deep learning models**

Differed from traditional classifiers that make predictions on individual examples, the long short-term memory (LSTM) network is capable of propagating information in a data sequence to improve prediction capacity. Literature studies have shown the effectiveness of LSTM-based models in processing sequence data such as free-text narratives and time-series signals. To achieve the best detection capacity, we also developed a LSTM-based model to capture growth trends and identify weight abnormalities. Weight feature sequence of each chart was used as model input. Zero padding was implemented to ensure that all charts had the same length of weight records. The padding values would be identified by the masking technique and ignored during model training and evaluation. A bi-directional LSTM (Bi-LSTM) was developed to aggregated information from a weight chart forwards and backwards to detect weight abnormalities. To simulate a prospective setting where future weights are not available, we also developed a one-direction LSTM model in the study. Appropriate variants such as LSTM with conditional random field models were explored but were excluded from further analysis due to lack of performance improvement.
Baseline approaches
The machine learning and deep learning models were compared with three baselines proposed in the literature. The first method was developed by CDC (denoted by CDC), which computed the modified weight-for-age z-scores based on a weight chart and considered weight points outside the range of [-5,8] normal. The second method was a computerized approach developed by Children’s Hospital of Philadelphia (denoted by CHOP). The method compared standard deviations of weight points against a weighted moving average in the chart to identify abnormal points that had significant deviation between recorded and expected values. We re-coded the algorithm based on the supplement code from the publication. Two error types identified by the CHOP method including 'duplicate' measurements on the same day and 'carried forward' weights within 90 days of prior measurement, were allowed and considered ‘normal’ in our study. The third method was a regression approach developed in our earlier study (denoted by REG), which modeled weight trend based on age, sex, previous weight values, and time from previous weight points to determine if a weight point was an outlier. It is worth noting that the CDC method identified weight abnormality based on the current weight point and its immediate previous one, while the CHOP and REG methods used all weight points in a chart.

Experimental setup
A stratified random sampling was performed based on individual subjects to split the MAIN dataset into two parts, 70% for training and 30% for evaluation. Ten-fold cross-validation was utilized to train the machine learning and deep learning models, where grid search parameterization was applied to optimize hyper-parameters including 1) cost parameters for L1- and L2-normalized LR, SVM-L, SVM-P and SVM-R (screened from 1e-6 to 1e6); 2) minimum number of observations in a node (3, 5, 10, and 20) and complexity parameters (screened from 1e-6 to 1e3, 0.3, 0.5 and 0.8) for DT; 3) number of trees (screened from 2^6 to 2^11) and minimum number of observations in a node (3, 5, 10, 15, 25, 40 and 50) for RFs; 4) optimal degree for SVM-P (2 and 3); 5) parameter γ for SVM-R (screened from 2^{-8} to 2^{-2}); 6) number of neurons (10, 15 and 20) and activation functions (rectified linear units [ReLU] and hyperbolic tangent [tanh]) for aNNs; 7) number of neurons (screened at 50 increments from 50 to 200) for LSTM and Bi-LSTM, and 8) learning rates for aNN, LSTM, and Bi-LSTM (screened from 1e-4 to 1e-2, 2e-4 and 5e-4). Stratified down-sampling was integrated into cross-validation to improve class imbalance. The machine learning and deep learning classifiers were implemented with R or Python.

Generalizability validation
To evaluate system generalizability, a separate set of 223,725 weight points were extracted for 7,190 patients in the CCHMC Discover Together Biobank. The data were collected as part of a larger study of the Electronic Medical Records and Genomics network. The dataset was processed using the same procedure as described above. Twenty percent of the charts were randomly selected and annotated by two study team members (Lei Liu and Dr. Ni) specialized in clinical informatics (denoted by the eMERGE dataset). The best-performing model as identified in our earlier experiments was trained on the full MAIN dataset and validated on the annotated eMERGE data. The results were compared with those generated from the baseline methods for model comparison.

Evaluation measures
We adopted the area under the ROC curve (AUC) as the primary measure and reported positive predictive value (PPV), sensitivity (SEN), negative predictive value (NPV), specificity (SPEC) when SEN reached 90% (a level required for production). To identify its limitations, we also performed an error analysis for the best-performing model on the MAIN test set, where patient charts with false positive or false negative predictions were visualized and inspected manually to identify potential causes of system errors.

Results
Descriptive statistics of the datasets
The summary of the annotation outcomes on the MAIN dataset (15,000 patient charts with 260,912 weight points) is presented in Table 1. We received a total of 775,670 annotation records. One hundred and forty-six records from 142 patient charts (0.95%; categories 3-4 in Table 1) contained display error and incomplete annotation and were excluded from analysis. The IRA between experts and the scoring-based outcome was substantial (Fleiss’ Kappa = 0.644, p-value <0.05). During data quality inspection we further excluded 14 charts (0.09%) due to clear abnormal weight values (i.e., LMS-based z-score ≥100), resulting in a set of 257,989 weight points from 14,844 patient charts. The abnormality rate of the dataset was 0.60% (1,549 abnormal weight values in 1,412 patient charts). After stratified sampling the training set contained 181,008 weight points from 10,391 charts, and the test set had 76,981 weight points from 4,453 charts. No missing values were presented in the training or test set. In the eMERGE dataset 215,736 weight points from 6,049 patients met inclusion criteria. We randomly selected 20% of the charts for annotation, resulting in a validation set of 44,747 weight points (1,209 charts) with an abnormality rate of 0.61%.
Table 1. Summary of the annotation outcomes on the MAIN dataset.

<table>
<thead>
<tr>
<th>ID</th>
<th>Description</th>
<th>Number of patient charts</th>
<th>Number of weight points</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Abnormal weight points</td>
<td>1,412</td>
<td>1,549</td>
</tr>
<tr>
<td>2</td>
<td>Normal weight entries</td>
<td>14,844</td>
<td>256,440</td>
</tr>
<tr>
<td>3</td>
<td>Charts containing display error in weight points</td>
<td>3</td>
<td>40</td>
</tr>
<tr>
<td>4</td>
<td>Charts containing incomplete annotations</td>
<td>139</td>
<td>2,364</td>
</tr>
<tr>
<td>5</td>
<td>Charts containing clear abnormal weight values (i.e., LMS-based z-score ≥100)</td>
<td>14</td>
<td>519</td>
</tr>
</tbody>
</table>

- Charts were excluded from analysis.

Predictive performance on the MAIN dataset

Figure 3 summarizes the performance of different machine learning models and baselines, where the detailed results are presented in Table 2. All machine learning and deep learning models achieved significantly better AUCs over the baseline approaches (p-value <0.05). Among the machine learning models, Bi-LSTM achieved the best performance, with an AUC of 0.996/0.989 on the training/test data. It also achieved the best PPV and SPEC when SEN was adjusted to ≥90%. The improvements of Bi-LSTM over the other models were statistically significant at 0.05 level. RF achieved the second-best performance, followed by aNN, SVM-R, and LSTM.

![AUC Graph](image)

Figure 3. Model performance for detection abnormal weight values. AUC: Area under the ROC Curve.

Table 2. Model performance for predicting abnormal weight values when SEN reached 90%.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>PPV</th>
<th>SEN</th>
<th>NPV</th>
<th>SPEC</th>
<th>AUC</th>
<th>PPV</th>
<th>SEN</th>
<th>NPV</th>
<th>SPEC</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR</td>
<td>1.82%</td>
<td>90.32%</td>
<td>99.87%</td>
<td>58.87%</td>
<td>0.885</td>
<td>1.90%</td>
<td>90.11%</td>
<td>99.91%</td>
<td>71.15%</td>
<td>0.909</td>
</tr>
<tr>
<td>SVM-L</td>
<td>1.48%</td>
<td>90.50%</td>
<td>99.89%</td>
<td>55.15%</td>
<td>0.904</td>
<td>2.23%</td>
<td>90.11%</td>
<td>99.92%</td>
<td>75.51%</td>
<td>0.924</td>
</tr>
<tr>
<td>SVM-P</td>
<td>2.36%</td>
<td>90.50%</td>
<td>99.92%</td>
<td>73.56%</td>
<td>0.922</td>
<td>2.35%</td>
<td>90.11%</td>
<td>99.92%</td>
<td>76.71%</td>
<td>0.918</td>
</tr>
<tr>
<td>SVM-R</td>
<td>2.76%</td>
<td>90.50%</td>
<td>99.93%</td>
<td>79.38%</td>
<td>0.944</td>
<td>3.14%</td>
<td>90.11%</td>
<td>99.93%</td>
<td>82.75%</td>
<td>0.942</td>
</tr>
<tr>
<td>DT</td>
<td>1.86%</td>
<td>95.29%</td>
<td>91.49%</td>
<td>48.09%</td>
<td>0.907</td>
<td>0.61%</td>
<td>99.16%</td>
<td>60.00%</td>
<td>0.01%</td>
<td>0.896</td>
</tr>
<tr>
<td>RF</td>
<td>4.46%</td>
<td>90.32%</td>
<td>99.93%</td>
<td>86.73%</td>
<td>0.961</td>
<td>4.66%</td>
<td>90.11%</td>
<td>99.93%</td>
<td>88.56%</td>
<td>0.961</td>
</tr>
<tr>
<td>aNN</td>
<td>4.54%</td>
<td>90.50%</td>
<td>99.93%</td>
<td>87.21%</td>
<td>0.960</td>
<td>3.95%</td>
<td>90.11%</td>
<td>99.93%</td>
<td>86.40%</td>
<td>0.957</td>
</tr>
<tr>
<td>LSTM</td>
<td>3.07%</td>
<td>90.06%</td>
<td>99.93%</td>
<td>82.87%</td>
<td>0.941</td>
<td>3.47%</td>
<td>90.11%</td>
<td>99.93%</td>
<td>84.41%</td>
<td>0.938</td>
</tr>
<tr>
<td>Bi-LSTM</td>
<td>41.82%</td>
<td>90.05%</td>
<td>99.94%</td>
<td>99.24%</td>
<td>0.996</td>
<td>35.91%</td>
<td>90.11%</td>
<td>99.94%</td>
<td>99.00%</td>
<td>0.989</td>
</tr>
<tr>
<td>BaseLine</td>
<td>PPV</td>
<td>SEN</td>
<td>NPV</td>
<td>SPEC</td>
<td>AUC</td>
<td>PPV</td>
<td>SEN</td>
<td>NPV</td>
<td>SPEC</td>
<td>AUC</td>
</tr>
<tr>
<td>CDC</td>
<td>0.65%</td>
<td>90.48%</td>
<td>99.43%</td>
<td>11.48%</td>
<td>0.553</td>
<td>0.52%</td>
<td>90.08%</td>
<td>99.54%</td>
<td>11.20%</td>
<td>0.546</td>
</tr>
<tr>
<td>REG</td>
<td>0.66%</td>
<td>90.48%</td>
<td>99.46%</td>
<td>12.22%</td>
<td>0.580</td>
<td>0.52%</td>
<td>90.08%</td>
<td>99.57%</td>
<td>12.07%</td>
<td>0.578</td>
</tr>
<tr>
<td>CHOP</td>
<td>0.62%</td>
<td>90.32%</td>
<td>99.55%</td>
<td>13.81%</td>
<td>0.618</td>
<td>0.64%</td>
<td>90.11%</td>
<td>99.53%</td>
<td>13.30%</td>
<td>0.620</td>
</tr>
</tbody>
</table>
Predictive performance on the eMERGE dataset
To assess its generalizability, the Bi-LSTM model optimized on the full MAIN dataset was compared with the baseline approaches on the eMERGE dataset (Table 3). We presented results with three probability thresholds: 1) the natural threshold (0.5), 2) the threshold when Bi-LSTM reached 90% SEN on the MAIN dataset (0.026), and 3) the threshold when Bi-LSTM reached 90% SEN on the eMERGE dataset (0.0014). Again, Bi-LSTM achieved significantly better performance than the baselines (p-value <0.05). Its AUC (0.989) was close to that achieved on the MAIN dataset (0.996/0.989 on training/testing). When using the probability threshold estimated by the MAIN dataset, the model had a SEN of 63.60% and a PPV of 68.65%. To yield a SEN of 90% on the eMERGE dataset, the system required a substantially lower threshold.

Table 3. Performance of the optimized Bi-LSTM model and three baselines on the eMERGE dataset.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>PPV</th>
<th>SEN</th>
<th>NPV</th>
<th>SPEC</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bi-LSTM (0.50)</td>
<td>92.22%</td>
<td>30.51%</td>
<td>99.58%</td>
<td>99.98%</td>
<td>0.989</td>
</tr>
<tr>
<td>Bi-LSTM (0.026*)</td>
<td>68.65%</td>
<td>63.60%</td>
<td>99.78%</td>
<td>99.82%</td>
<td>0.989</td>
</tr>
<tr>
<td>Bi-LSTM (0.0014**)</td>
<td>16.97%</td>
<td>90.07%</td>
<td>99.94%</td>
<td>97.30%</td>
<td>0.989</td>
</tr>
<tr>
<td>Baseline</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CDC</td>
<td>0.61%</td>
<td>90.07%</td>
<td>99.40%</td>
<td>10.44%</td>
<td>0.510</td>
</tr>
<tr>
<td>REG</td>
<td>0.62%</td>
<td>90.07%</td>
<td>99.49%</td>
<td>12.33%</td>
<td>0.586</td>
</tr>
<tr>
<td>CHOP</td>
<td>0.63%</td>
<td>90.07%</td>
<td>99.52%</td>
<td>13.12%</td>
<td>0.605</td>
</tr>
</tbody>
</table>

*The probability threshold when Bi-LSTM reached 90% SEN on the MAIN dataset. **The probability threshold when Bi-LSTM reached 90% SEN on the eMERGE dataset.

Error analysis
By adjusting SEN to 90.11%, the optimized Bi-LSTM model achieved a PPV of 35.91% on the MAIN test set, resulting in 764 false positive (in 617 charts) and 47 false negative predictions (in 44 charts). We performed an error analysis on these patient charts to identify potential causes of error. Table 4 summarizes the error categories and the example charts are visualized in Figure 4.

Table 4. Categorization and distribution of false positives (a) and false negatives (b) made by Bi-LSTM on the MAIN test set.

<table>
<thead>
<tr>
<th>ID</th>
<th>Category description</th>
<th>False positives</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Weight fluctuation between multiple measurements within 48 hours.</td>
<td>196</td>
<td>25.65%</td>
</tr>
<tr>
<td>2</td>
<td>A patient had a large weight increase/decrease because of 1) fast weight loss or weight gain, 2) frequent weight fluctuation, 3) weight gain above 97th percentile or weight loss below 3rd percentile and 4) that the previous or the next weight point was labeled as ‘abnormal’ by annotators.</td>
<td>183</td>
<td>23.95%</td>
</tr>
<tr>
<td>3</td>
<td>The current point was the first/last point in a chart and lack past or future information.</td>
<td>176</td>
<td>23.04%</td>
</tr>
<tr>
<td>4</td>
<td>Experts and the model had different tolerance to weight changes at data points with low weight values.</td>
<td>167</td>
<td>21.86%</td>
</tr>
<tr>
<td>5</td>
<td>Errors with unidentified reasons</td>
<td>42</td>
<td>5.50%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ID</th>
<th>Category description</th>
<th>False negatives</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>There was a large weight change between two adjacent measurements with a long time interval.</td>
<td>23</td>
<td>48.94%</td>
</tr>
<tr>
<td>2</td>
<td>The current point was the first/last point in a chart and lack past or future information.</td>
<td>9</td>
<td>19.15%</td>
</tr>
<tr>
<td>3</td>
<td>Experts and the model had different tolerance to weight changes at data points with low weight values.</td>
<td>7</td>
<td>14.89%</td>
</tr>
<tr>
<td>4</td>
<td>Weight fluctuation between multiple measurements within 48 hours.</td>
<td>4</td>
<td>8.51%</td>
</tr>
<tr>
<td>5</td>
<td>Two adjacent data points had similar age and weight values and were both annotated as ‘abnormal’ by annotators, while the classifier only predicted one as ‘abnormal’.</td>
<td>4</td>
<td>8.51%</td>
</tr>
</tbody>
</table>
Figure 4. Example errors made by the optimized Bi-LSTM model on the MAIN test set. Arrows indicate the false positive/false negative cases in the error analysis.
Discussion
In this study we developed an AWAD to analyze patient weight charts and detect abnormal values from a pediatric population and setting. Compared with the error detection methods proposed in literature, the machine learning-based approaches demonstrated significantly better detection capacity (Table 2). The finding illustrates the advantage of machine learning technologies over knowledge-driven rules by learning latent patterns from the data. In particular, the deep learning-based Bi-LSTM achieved excellent performance and its improvements over the other classifiers were statistically significant (p-value <0.05). The PPV (35.91%) and SEN (90.11%) achieved by Bi-LSTM on the test set suggested that when using AWAD, a researcher could capture over 90% of abnormal weight values by reviewing only 1.56% of the data. Nevertheless, the one direction LSTM did not show improved performance over traditional classifiers such as RF, aNN and SVM-R, suggesting that the advantage gained by bi-LSTM was the aggregated information from future weight points. The similar performance of Bi-LSTM on the eMERGE dataset confirmed its improvements over the literature methods and validated its generalizability (Table 3). The results also suggested that different probability thresholds would be needed to balance sensitivity and specificity on different datasets.

The developed algorithms and findings could have potential for a significant impact on both research and clinical care. For instance, the Bi-LSTM approach can be used to facilitate data cleaning and improve data quality for secondary analysis (e.g., EHR-based phenotyping as conducted by the eMERGE network). Its PPV and SEN (35.91%/90.11%) suggest potential for more than 98% screening effort reduction in weight abnormality detection. In practice, the abnormality predictions could be enumerated with an empirical probability threshold to balance sensitivity and specificity. On the other hand, the developed system can be implemented in prospective settings to improve healthcare delivery. For example, weight-based medication orders are created based on the most recent weight point in EHRs. By predicting and visualizing abnormality predictions on a weight chart, additional review and timely correction could be made to mitigate effects of entry errors on patient safety. Even though Bi-LSTM is not directly applicable due to unavailability of future weight information, the high performance achieved by the second-best RF (0.961/0.961 AUCs on the MAIN training/test sets) still assures the effectiveness of our application.

Error analysis, limitations and future work
The error analysis on the Bi-LSTM uncovered several areas of improvement. Over 25% of false positives (category 1 in Table 4a) were due to weight fluctuation in measurements within a very short time period (48 hours). Similarly, a large portion of errors (category 4) were caused by different tolerance to change between the annotations and model predictions when weight values were small, particularly at younger ages. We hypothesized that the annotators could have different tolerance of weight variation in a short time period and on low-weight points such that the model captured a “compromised” tolerance that optimized the predictive performance. This issue could potentially be addressed when we perform the study in prospective settings, where the weight abnormalities collected are true weight errors. In addition, the model often labeled large weight changes as abnormal (category 2), which however, could be weight gain or weight loss caused by factors such as health conditions, medication effects, and changes in living environment. Enriching the feature set with more comprehensive patient information (e.g., medical history, medication use) may contribute to better performance, and should be explored in future studies. Finally, 23% of errors (category 3) were due to lack of information because the weight point was the first or last one in a chart. The finding revealed a major limitation of data-driven technologies and incorporating knowledge-based rules might help mitigate this issue. For false negatives, approximately 50% of errors (category 1 in Table 4b) were caused by a pattern when there was a large weight change between two adjacent measures with a long time interval. The errors might be caused by insufficient documentation between ages and will be further investigated in the future. The other causes (categories 2-4) were similar to those from false positives. Few errors were caused by successive abnormal weight points (category 5), which could potentially be addressed by data post-processing (e.g., if an abnormal point is detected, any adjacent points with close weight and age are classified as abnormal).

While this study makes an important contribution to advancing methods for weight abnormality detection, there are limitations to be considered. First, the analysis excluded weight measures before age two that could have more complex weight changing patterns. Additional development and evaluation are therefore required for this age group. Likewise, the system performance is based on data from a single pediatric institution. Built upon the work, a web-service-based AWAD has been developed to analyze weight charts, identify abnormal weight values, and visualize weight trends and errors. Once the web-service-based AWAD has been deployed, we anticipate assessing its generalizability with more diverse patient populations and institutions. It is worth noting that machine learning techniques support retraining the model when new data become available. As such, if generalizability is not satisfactory, appropriate active learning approaches could be implemented to re-tune the system automatically as new data become available7. Finally, this study was restricted to retrospective data. Once reliability and generalizability are established, the AWAD
can be transferred to a production environment to adequately assess its usability and utility with prospective data in future studies.

Conclusions
By utilizing machine learning technologies, we developed an automated approach, AWAD, to analyze pediatric weight charts and identify abnormal weight values. In two reference-standard based evaluation of real-world clinical data, the machine learning models showed good capacity for detecting weight abnormalities and they significantly outperformed the methods proposed in literature. The best-performing model (Bi-LSTM) achieved AUCs ≥0.989 across the two datasets, with good PPVs when sensitivities were adjusted to ≥90%. Given its high performance in this stage of development, we hypothesize that the AWAD, when fully deployed, holds great potential to facilitate clinical research and healthcare delivery that rely on accurate and reliable weight measures.

Acknowledgements
We thank Mr. Milan Parikh at University of Cincinnati for his effort on reviewing and documenting the scripts of the baseline approaches. We also thank the eMERGE network for providing the weight dataset. Particular thanks go to PJ Van Camp who developed the visual annotation tool and managed the annotation process.

References
De-identifying Socioeconomic Data at the Census Tract Level for Medical Research Through Constraint-based Clustering

Yongtai Liu, MS1, Douglas Conway, BA2, Zhiyu Wan, PhD1, Murat Kantarcioglu, PhD3, Yevgeniy Vorobeychik, PhD4, Bradley A. Malin, PhD1,2
1Vanderbilt University, Nashville, TN; 2Vanderbilt University Medical Center, Nashville, TN; 3University of Texas at Dallas, Richardson, Texas; 4Washington University in St. Louis, St. Louis, MO

Abstract

Numerous studies have shown that a person’s health status is closely related to their socioeconomic status. It is evident that incorporating socioeconomic data associated with a patient’s geographic area of residence into clinical datasets will promote medical research. However, most socioeconomic variables are unique in combination and are affiliated with small geographical regions (e.g., census tracts) that are often associated with less than 20,000 people. Thus, sharing such tract-level data can violate the Safe Harbor implementation of de-identification under the Health Insurance Portability and Accountability Act of 1996 (HIPAA). In this paper, we introduce a constraint-based k-means clustering approach to generate census tract-level socioeconomic data that is de-identification compliant. Our experimental analysis with data from the American Community Survey illustrates that the approach generates a protected dataset with high similarity to the unaltered values, and achieves a substantially better data utility than the HIPAA Safe Harbor recommendation of 3-digit ZIP code.

Introduction

Numerous studies have shown that a person’s health status is associated with their socioeconomic status (SES)1,4. For instance, in some populations, people with higher education attainment, larger income, and more prestigious occupations enjoy better health and longer lives5. By contrast, a lower level of socioeconomic status is associated with an increased risk of chronic disorders, such as cardiovascular disease6,8 and mortality9. It is evident that incorporating a patient’s SES information into clinical records, and research datasets, will promote more meaningful, as well as potentially actionable, medical investigations10.

However, in many situations, a patient’s SES information is not readily available11. In lieu of this information, it is possible to estimate a person’s SES level based on the neighborhood in which they live. Notably, the SES for a small geographic area, in the form of a U.S. Census tract, can be derived from the 5-year estimates of the American Community Survey (ACS)12 - information that is publicly available online. For the purposes of our analysis, we prioritized the set of census derived variables from Brokamp et al.13. An example is shown in Table 1. The socioeconomic factors in the dataset include: 1) assisted income, 2) high school education level, 3) median household income, 4) lacking health insurance level, 5) poverty level, 6) vacant housing rate, and 7) a deprivation index, which is a single value assigned to the census tract that provides a numerical summarization of the area’s overall deprivation. The deprivation index ranges from 0 to 1, with a higher value indicating a larger level of SES hardship. It was created based on a principal component analysis (PCA)14 on the six aforementioned socioeconomic factors in the ACS dataset. Throughout this paper, we refer to the six ACS factors and the deprivation index as census tract socioeconomic factors.

It is often the case that electronic health records (EHRs), and affiliated information, are utilized in a de-identified manner, as defined by the Health Insurance Portability and Accountability Act of 1996 (HIPAA). While census tract socioeconomic factors have been relied upon to estimate a patient’s SES in various research studies, such information has not been commonly integrated into de-identified records. One of the primary reasons for such a lack of SES information is that EHR data is often de-identified using the Safe Harbor implementation of the HIPAA de-identification standard. Under Safe Harbor, no geographic information granular than the initial three digits of the ZIP code (or a comparable indicator of geography) may be disclosed if it contains less than 20,000 people15. Census tracts tend to be geographic regions with relatively small populations. Specifically, the population covered by a census tract ranges from 20 to 65,000 people, with an average of 4,000 people. This is problematic because SES data elements are often unique to census tracts. For example, in 2017, only one census tract in the country has a median income of

---

1 Available for download from https://www.census.gov/data/developers/data-sets/acs-5year.html
entities be used to satisfy the Privacy Rule’s determination method requires that every census tract group covers at least 20,000 inhabitants. We achieved this by first clustering all the tracts into small groups, then iteratively merging groups that fail to satisfy the privacy requirements to their most similar group until all groups meet the requirements. We assessed the utility of the de-identified data with five years of ACS socioeconomic datasets. The result suggests that our approach achieves a substantially better data utility than HIPAA Safe Harbor recommended 3-digit ZIP code grouping. We also found that k-means model generates data with the highest overall utility, based on six socioeconomic factors and the composite deprivation index.

Table 1. Examples of six socioeconomic factors derived from the ACS dataset and the deprivation index. The Census Tract FIPS code is an 11-digit number that uniquely identifies each census tract in the United States. Rows 1 and 4 show two census tracts and their population, original SES values before de-identification. Rows 2 and 5 show the de-identification result of the two census tracts based on 3-digit ZIP code grouping. Rows 3 and 6 are the corresponding de-identification result based on our constraint-based clustering (CBC) approach.

<table>
<thead>
<tr>
<th>Census Tract FIPS</th>
<th>De-id method</th>
<th>Population</th>
<th>Percentage Assisted Income</th>
<th>Percentage High School Education</th>
<th>Median Household Income</th>
<th>Percentage Lacking Health Insurance</th>
<th>Percentage Poverty</th>
<th>Percentage Vacant Housing</th>
<th>Deprivation Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>19089-960100</td>
<td>3-digit ZIP (521**)</td>
<td>3,884</td>
<td>12.12%</td>
<td>91.4%</td>
<td>$47,000</td>
<td>4.4%</td>
<td>13.32%</td>
<td>7.46%</td>
<td>0.30</td>
</tr>
<tr>
<td>CBC</td>
<td></td>
<td>58,083</td>
<td>8.5%</td>
<td>92.0%</td>
<td>$51,608</td>
<td>4.2%</td>
<td>10.70%</td>
<td>8.33%</td>
<td>0.30</td>
</tr>
<tr>
<td>39143-961300</td>
<td>3-digit ZIP (434**)</td>
<td>4,069</td>
<td>12.64%</td>
<td>92.5%</td>
<td>$50,553</td>
<td>4.6%</td>
<td>13.77%</td>
<td>7.27%</td>
<td>0.29</td>
</tr>
<tr>
<td>CBC</td>
<td></td>
<td>158,817</td>
<td>15.50%</td>
<td>88.6%</td>
<td>$46,495</td>
<td>8.5%</td>
<td>16.47%</td>
<td>8.60%</td>
<td>0.37</td>
</tr>
<tr>
<td></td>
<td></td>
<td>22,835</td>
<td>12.01%</td>
<td>91.6%</td>
<td>$48,304</td>
<td>4.3%</td>
<td>12.71%</td>
<td>7.26%</td>
<td>0.29</td>
</tr>
</tbody>
</table>

Related Work

**HIPAA Safe Harbor**

The U.S. Congress enacted HIPAA to standardize the sharing of health insurance claims, as well as provide common protections for sensitive patient health information from being disclosed without a patient’s consent or acknowledgment. Several years later, the U.S. Department of Health and Human Services (HHS) issued the HIPAA Privacy Rule to further elaborate on privacy requirements. At the same time, HHS created the notion of a de-identified dataset, which corresponds to data that is no longer protected by the regulation, as it does not directly, and is unlikely to indirectly, uniquely identify the corresponding patient. The regulation indicates that two alternative methods can be used to satisfy the Privacy Rule’s de-identification standard: Expert Determination and Safe Harbor. Covered entities (and their business associates) must ensure data is compliant with one of the methods. The Expert Determination method requires that an expert use generally accepted statistical and/or scientific methods to confirm
that the chance a record could be re-identified is very small. Statistical perturbation methods, such as differential privacy\textsuperscript{18}, may provide support to de-identify aggregated geographic data for biomedical research.

By contrast, Safe Harbor provides a more straightforward approach to de-identification. Specifically, it enumerates 18 types of identifiers that must be removed before covered entities (or their business associates) can claim the data is compliant with de-identification. As noted, geographic location is one such type of information. Specifically, all geographic subdivisions smaller than a state must be removed, except for the initial three digits of the ZIP code if combining all ZIP codes with the same three initial digits contains more than 20,000 people. Thus, to devise a method that is compliant with Safe Harbor, it is necessary to ensure that each de-identified record covers multiple census tracts such that the combining census tracts contain more than 20,000 people.

\textit{K-anonymity}

K-anonymity is a privacy protection model designed to ensure data is non-unique\textsuperscript{16}. Specifically, a dataset is \textit{k}-anonymous if each record is indistinguishable from at least \textit{k}-1 other records in the same dataset. One the most popular approaches to achieve \textit{k}-anonymity is through generalization and suppression of data. Generalization replaces a data attribute's actual value with a more general value (or a range of values). Often this is accomplished via the assistance of value generalization hierarchy. Using the ZIP code as an example, it can be seen that 37203 and 37206 can be generalized to 3720\textsuperscript{6}. By contrast, suppression removes an entire attribute (or an outlying record) from the dataset. Achieving \textit{k}-anonymization is known to be an NP-hard problem\textsuperscript{13}, but there are several effective approximation algorithms that have been developed\textsuperscript{19-20}. As we show below, with some modification, these algorithms can be adapted to solve our problem.

\textit{Constraint-based Clustering}

Constraint-based clustering aims to find a clustering for data while satisfying a set of constraints. However, methods developed to date\textsuperscript{21,22} focus on constraints for specific data instances. For instance, a method might include certain \textit{must-link} constraints or \textit{cannot-link} constraints between data points. In our problem, there are no pre-defined constraints for the data points. Rather, the privacy constraints correspond to the clusters themselves.

Generally, the constraint-based optimization task can be modeled as an Integer Linear Programming (ILP)\textsuperscript{3} problem, where the privacy requirements can be addressed as constraints and the goal is to minimize the intra-cluster distance. Unfortunately, our problem is known to be NP-complete and, given the large search space (more than 70,000 census tracts by more than 10,000 clusters), such a program cannot find an optimal solution in a realistic amount of time.

\textbf{Materials and Methods}

In this study, we utilize the 2017 census tract-level deprivation index dataset\textsuperscript{4}. From the ACS data, we also generate four years of census tract-level deprivation index datasets from 2012 to 2015, according to Brokamp et al.\textsuperscript{2}. These four datasets have the same structure as the 2017 dataset. Due to a documentation problem, we did not generate a census tract-level deprivation index dataset for the 2016 ACS data. The 2017 dataset contains 72,943 census tracts, listed by their federal information processing standards (FIPS) ID and corresponding deprivation index, along with six ACS features used to derive the deprivation index. The population of each census tract can be acquired from the U.S. Census Bureau website\textsuperscript{5}. We represent the dataset with six ACS features as a matrix \textit{X}, where \textit{X\textsubscript{i}} represents row \textit{i}, a census tract record, \textit{X\textsubscript{j}} represents the \textit{j}th column, an ACS feature. The deprivation index is denoted by \textit{y}, where \textit{y\textsubscript{i}} represents the deprivation value of census tract \textit{t\textsubscript{i}}. We represent the population of tract \textit{t\textsubscript{i}} as \textit{pop\textsubscript{i}}.

In this section, we first introduce the de-identification pipeline and several clustering methods considered in this investigation. Next, we present the evaluation criteria to select the best clustering methods. As we indicated earlier, the unique linkage between a census tract ID and the values of socioeconomic factors for a census tract may induce the disclosure of a person’s geolocation. We mitigate this risk by changing the mapping between census tract and socioeconomic factor values from \textit{a one-to-one} mapping to \textit{a many-to-one} mapping. This is accomplished by grouping census tracts into clusters based on their similarities and replacing them with their mean socioeconomic factor values for census tracts associated with each cluster. This ensures that, one socioeconomic factor value is guaranteed to

\footnotesize
\begin{itemize}
\end{itemize}

795
correspond to multiple census tracts and, thus, encompasses all of the population in the tracts. To comply with the HIPAA Safe Harbor requirements, each cluster must satisfy the following privacy requirements:

1. Each cluster must contain at least two census tracts. \((\text{MinTract} > 2)\)
2. Each cluster must cover at least 20,000 individuals. \((\text{MinPop} > 20,000)\)

The de-identification pipeline is depicted in Figure 1. Given the dataset with six ACS features, we apply a clustering method. We considered several clustering methods to determine which would be most appropriate for the task at hand. Specifically, we investigated hierarchical models in the form of Agglomerative Clustering\(^2\) and divisive \(k\)-means\(^3\), a density based clustering model DBSCAN\(^4\), the classic \(k\)-means model\(^5\), and a Gaussian Mixture Model\(^6\), which is in effect a generalized \(k\)-means model. Some of these models require certain parameters to be defined by the user, such as the number of clusters in \(k\)-means. To determine an appropriate value for such a setting, we experimentally analyzed a range of values.

We also recognized that certain socioeconomic studies prioritize deprivation over all of the ACS features. As such, we also considered a greedy clustering methodology, that is based solely on the deprivation index (neglecting the six ACS features).

Given that the clustered candidate datasets are not guaranteed to satisfy the privacy requirements, when necessary, we apply a post-processing step to process such clusters to satisfy the \(\text{MinTract}\) and \(\text{MinPop}\) conditions. After completion of this post-processing step, we calculate the new deprivation index, which corresponds to the first principle component of the clustered dataset. We select the best clustering in the evaluation step and output the de-identified census tract-level socioeconomic dataset.

In the following subsections, we will introduce the selected clustering methods, the post-processing algorithm, the evaluation criteria, and a verification experiment with different years of the ACS data.

![Figure 1.](image)

**Figure 1.** The de-identification pipeline for census tract-level socioeconomic dataset.

**Clustering Models**

**Greedy Algorithm**

This clustering approach is based purely on the deprivation index and neglects the ACS features. It initially sorts the deprivation value in ascending order and then groups adjacent tracts into the current cluster (from top to bottom). Once the current cluster has a sufficient number of census tracts to satisfy the privacy requirements (i.e., \(\text{MinTract} > 2\) and \(\text{MinPop} > 20,000\)), this cluster is considered complete. The algorithm then generates a new cluster. This process iterates until all census tracts are clustered. The new deprivation index \(\bar{y}\) and the new ACS features \(\bar{X}\) are then the mean of the corresponding values \(y\) and \(X\) in the cluster. Since the clusters generated by the Greedy algorithm already satisfy the privacy requirements, this algorithm does not require any post-processing.

Unlike the Greedy algorithm, the following approaches focus on \(X\) as the dataset to be clustered.

**Hierarchical (Agglomerative) Clustering**

This algorithm begins with each record \(X_i\) as a singleton cluster. Next, the algorithm iteratively merges the two most similar clusters until there is only one cluster (i.e., the entire dataset) left. The similarity is measured as the sum of squared distances of all census tracts of pairs of clusters. After clustering, the algorithm walks down the generated hierarchical tree from top to bottom to find all the relevant clusters. Specifically, a node is a cluster of interest if, and only if, 1) the node satisfies the two privacy requirements and 2) at least one of its child nodes fails to satisfy the requirements. Note that this approach does not require any post-processing. The new census tract record \(\bar{X}_i\) is the mean of census tracts in one cluster and the new deprivation index \(\bar{y}\) is computed based on \(\bar{X}\).
**Divisive k-means**

This is a top-down hierarchical clustering approach. We initiate the method by assigning all census tracts to a single cluster. We then iteratively select the divideable cluster with the most population from the pool and split it into two new clusters using 2-means. The algorithm stops when there are no clusters can be split.

**K-means**

In this approach, we use the classic k-means to cluster the dataset into k clusters. k-means aims to make the intra-cluster points as similar as possible through an Expectation-Maximization (EM) process. In the E-step, the algorithm assigns points to their nearest cluster, according to the Euclidean distance from the point to the cluster centroid. In the M-step, the algorithm updates the centroids of each cluster. This process iterates until the system is stable (or until a maximum number of iterations is reached). To maximize the similarity between the clustering and the original data, the size of each cluster should be as small as possible, which implies that the choice of k should be relatively large. In our experiments, we vary k from 14,000 to 70,000, with a step size of 4,000. It is possible that some clusters generated by k-means fail to meet the privacy requirements, such that we perform post-processing after clustering.

**DBSCAN**

This is a density-based clustering model that assumes that a cluster should be determined by the density of data distribution. DBSCAN measures the density and extends a cluster based on reachability. It uses neighborhood parameters r (radius of a neighborhood) and MinPts (minimum points within a distance of r) to describe the reachability. Based on the data, we plot the k-distance graph, and set MinPts to 3 and r to 0.5. DBSCAN returns a clustering with noise (data points do not belong to any clusters). To ensure the clustering result meets the requirements, we treat each noise point as a new cluster and perform post-processing accordingly.

**Gaussian Mixture Model**

This is a probabilistic model that assumes data points are generated and sampled from a mixture of k Gaussians with unknown parameters (means and covariance of Gaussians). The cluster of each data point is determined by the posterior probability of mixture components given the point. GMM can be solved by an iterative EM algorithm. The E-step calculates the posterior probability of the data points (i.e., census tracts) belonging to a cluster based on the parameters of the latest iteration. In the M-step, the algorithm updates parameters using maximum likelihood estimation given current clustering. In our experimental analysis, we vary the number of components (k) from 4,000 to 8,000.

**Post-Processing**

To post-process a clustering, we iteratively modify the clusters that fail the privacy requirements until no such clusters are left. In each iteration, the algorithm selects the cluster with the smallest population and reassigns the corresponding census tracts to their nearest cluster, according to the Euclidean distance between tracts and cluster centroids. The algorithm then updates all of the clusters' centroids. This process iterates until no clusters require post-processing.

**Evaluation Criteria**

Finally, we select the model that is most similar to the original dataset. We use KL-divergence to measure the similarity between the de-identified dataset and the original data. KL-divergence is an information-theoretic approach that measures the statistical distance between two distributions. Specifically, given two probability distribution P and Q on the same probability space, X, where P is the true (i.e., observed) distribution and Q is the approximated (i.e., modeled) distribution, the KL-divergence from Q to P, $D_{KL}(P||Q)$, is defined as the following equation:

$$D_{KL}(P||Q) = \sum_{x \in X} P(x) \log \left( \frac{Q(x)}{P(x)} \right)$$

The divergence score is non-negative, and a score of 0 indicates the two distributions are identical, where a higher divergence score suggests a larger divergence from Q to P. It should be recognized that, unlike Jenson Shannon (JS) divergence, the KL-divergence is an asymmetric distance measure, such that $D_{KL}(P||Q) \neq D_{KL}(Q||P)$. The intuition is that when the true probability $p$ for an event $x$ is large, but the approximated probability $q$ is small, there is a larger divergence. By contrast, when the approximated probability $q$ is large but the true probability $p$ is small, the divergence is not as large as the previous case.
We adopt KL-divergence instead of other symmetric measures because it is a natural fit for our problem. The original dataset is the true distribution of the data, and the de-identified protected datasets can be treated as the approximated distributions. We are only concerned about the projection from the original into the protected space, instead of the inversion. Thus, we compute the KL-divergence for all features and compare the models in two scores: the KL-divergence of the deprivation index and the sum of the KL-divergence of all features. These two scores represent two different usage scenarios. That is, when a researcher is more interested in using only one feature to capture the entirety of community deprivation level, then the deprivation index divergence score should drive which clustering to select. And, when a researcher is concerned about the overall data utility, then overall KL-divergence should serve the selection criterion.

**Experimental design**

We apply our de-identification pipeline on the 2017 census tract-level deprivation index dataset and select the clustering model that yields the protected clustering with the lowest KL-divergence. Given that this process is based on one year’s data only and that this selection may suffer from certain bias, we assess the generalizability of our model selection. To do so, we run the same experiment on four additional census tract-level datasets, one from each of the years between 2012 and 2015.

**Results**

Figure 2 presents the KL-divergence scores of the protected datasets generated by the six clustering models. The y-axis represents the overall KL-divergence score of seven socioeconomic factors and the x-axis is the divergence score of the deprivation index. The lowest KL-divergence achieved by the Greedy algorithm (for the deprivation index) was only 0.018, but the overall divergence is substantially higher than the other models (4141.05). The k-means model achieved the smallest overall KL-divergence (388.96) and the second smallest KL-divergence for the deprivation index (7.92). We do not show the DBSCAN model in this figure because the divergence score is too high (overall 13136, deprivation index 1667). This method was discarded from further consideration given its obvious inferiority to the other clustering methods.

![Figure 2](image.png)

**Figure 2.** The KL-divergence of the protected datasets generated by the various clustering models. The y-axis represents the sum divergence score of seven socioeconomic features, whereas the x-axis is the KL-divergence score of the deprivation index. The six clustering models are: Greedy, hierarchical clustering (Agg), divisive k-means (DivKM), k-means, Gaussian mixture model (GMM), and 3-digit ZIP code (ZIP3). The results for DBSCAN model not shown as it is substantially higher in divergence than the other methods.

In addition, for comparison purposes, based on HIPAA Safe Harbor recommendation, we grouped the census tracts according to their 3-digit ZIP code, which we refer to as the ZIP3 de-identified dataset. It is evident that except for
DBSCAN, the ZIP3 model exhibits the highest overall divergence score (1414.27) and the highest deprivation index divergence score (414.11).

**Figure 3.** The change in the KL-divergence as a function of the initial $k$ in the $k$-means clustering process. The orange line corresponds to the sum of KL-divergence of the seven socioeconomic features. The purple line corresponds to the KL-divergence of the deprivation index.
Figure 4. The overall KL-divergence of the protected datasets generated by four clustering methods for the socioeconomic dataset associated with each of the years between 2012 and 2015. The four clustering models depicted correspond to hierarchical clustering (Agg), divisive $k$-means (DivKM), $k$-means, and the Gaussian mixture model (GMM).

Given the strong performance of the $k$-means clustering, we performed an in-depth analysis of this method. The $k$-means model in Figure 2 was initialized with 70,000 clusters. After post-processing, the model yields 10,429 clusters, with a median cluster size (the number of census tracts in one cluster) of 7. The change in the KL-divergence as a function of the initial $k$ is shown in Figure 3. It can be seen that the overall KL-divergence (shown in orange) drops quickly from an initial $k$ of 14,000 to 50,000. After this point, the change in the overall KL-divergence is relatively stable, but KL-divergence is minimized at a $k$ of 70,000, with a score of 390.72. The deprivation index line (shown in purple) shows similar behavior; however, the lowest deprivation divergence happens at a $k$ of 58,000, with a deprivation index divergence score of 7.81.

To assess the robustness of the clustering with the 2017 data, we compare the clustering models on the ACS datasets from each of the four years from 2012 through 2015. Figure 4 shows the overall KL-divergence for the projected datasets generated by the four clustering models in four years of data. It can be seen that the $k$-means model consistently achieves the smallest overall KL-divergence in all four years, followed by the hierarchical clustering, Gaussian Mixture Model, and the Divisive $k$-means.

Figure 5 shows the KL-divergence for the deprivation index for 2012 through 2015. In this situation, the Greedy algorithm achieves the lowest deprivation index KL-divergence (below 0.3) every year. The KL-divergence for the $k$-means model is around 8, while the other models all have scores greater than 10. Save for the Greedy algorithm, it can be seen that the $k$-means method generates the protected dataset with the lowest overall KL-divergence and the lowest deprivation index KL-divergence.

Discussion and Conclusions

This work introduced a pipeline to generate a geography-influenced socioeconomic data that maximizes its utility while remaining compliant with the de-identification requirements of the HIPAA Safe Harbor implementation standard. The pipeline consists of three parts: 1) clustering, 2) post-processing, and 3) evaluation. The pipeline can
automatically generate the protected dataset from a set of clustering models given predefined privacy requirements. We assessed the utility of the de-identification process with five years of ACS socioeconomic datasets. The result suggests that a $k$-means model generates data with the highest overall utility and the second-highest utility for the deprivation index. Notably, the de-identified dataset generated by the $k$-means model is substantially more similar to the original data than that achieved by grouping by the 3-digit ZIP code as suggested by the HIPAA Safe Harbor policy. We further showed that a simple greedy algorithm, which clusters data based on only the deprivation index, leads to the best utility for the deprivation index, but at a significant cost to the utility of other socioeconomic features.

For $k$-means clustering, when selecting a value of $k$, it is evident that larger values do not necessarily lead to better overall utility. Our empirical results illustrated that this was true both for the overall KL-divergence (for all socioeconomic features), as well as when the focus is on a single value (for the deprivation index). This suggests that one needs to consider a wide range of values when selecting the initial number of clusters $k$.

Finally, there are certain limitations to this work, which we believe serve as opportunities for future research. First, while KL-divergence is an information theoretic measure of the similarity of probability distributions, the values associated with census tract-level socioeconomic factors are not probability distributions. Second, though we evaluate data utility with a divergence score, this only illustrates similarity of the distribution of the data and does not indicate how the data changes influence any associations that may be inherent in the data.

**Acknowledgements**

This research was sponsored, in part, by NIH grants R01HG006844, RM1HG009034, U54MD010722 and UL1TR002243.

**References**

Application of the i-PARIHS framework in the implementation of speech recognition technology as a way of addressing documentation burden within a mental health context

Brian Lo, MHI1,2, Khaled Almilaji, MD, MHI1, Damian Jankowicz, PhD1, Lydia Sequeira, MHI1,2, Gillian Strudwick, RN, PhD1,2, Tania Tajirian, MD1,2
1Centre for Addiction and Mental Health, Toronto, Ontario, Canada; 2University of Toronto, Ontario, Canada

Abstract

Documentation burden continues to be a critical issue in the adoption of comprehensive electronic health record systems. This case study demonstrates how the i-PARIHS framework can be applied to support the implementation of interventions in reducing documentation and EHR-related burden in a mental health context. As part of pre-adoption implementation activities for Speech Recognition Technology (SRT), a cross-sectional survey was conducted with physicians, residents, and fellows at an academic mental health hospital to explore their perceptions on SRT. Open-ended responses and follow-up interviews explored challenges and concerns on using SRT in practice. Through an analysis using the i-PARIHS framework, key considerations were mapped across the four components of the framework. This study demonstrates the value of applying well-established implementation frameworks, such as the i-PARIHS framework, in mitigating challenges related to documentation burden. Future studies should explore how implementation frameworks can be systematically embedded in addressing EHR-related burden.

Introduction

As electronic health record (EHR) systems with advanced features continue to be adopted worldwide1, unintended consequences related to documentation burden and potentially clinician burnout are increasingly reported2-4. Studies examining clinician burnout have suggested that physicians spend approximately double the time on documentation when compared to time dedicated to direct patient care5. These challenges are not unique to a profession or clinical domain, and are said to jeopardize patient care and safety6. Recently, the National Academies of Sciences, Engineering, and Medicine3 have recognized this challenge in their latest report on clinician burnout and healthcare professional well-being. Thus, there is a need to address these challenges in a timely manner.

As part of our organization’s Physician Engagement Strategy, we conducted a baseline survey with physicians to examine their level of satisfaction and EHR-related burnout7. Of the 208 respondents, about one in four physicians reported experiencing burnout related to use of the EHR7. Given that documentation burden is a well-established factor of burnout8, we decided to implement speech recognition technology (SRT) to mitigate documentation burden and clinician burnout7. SRT allows clinicians to directly dictate into the EHR system and is expected to enhance quality of care through reduced turnaround time and more accurate reports9-11. While the use of SRT mostly originated in radiology, it is increasingly adopted across primary care and internal medicine disciplines12. With the narrative documentation style in psychiatry, we hypothesize that SRT will reduce documentation time and increase efficiency with using the EHR. However, there remains a gap in evidence and guidance on the implementation of SRT in mental health settings. Our organization has previously rolled out a local, outdated version of SRT to a small group of physicians, however, it was not optimized and had low adoption levels over time13,14. Therefore, it was an organizational priority to explore a more robust implementation approach for the upcoming site-wide implementation of the improved SRT. The improved SRT solution is a cloud-based solution that promises improved accuracy of the dictations as well as the use of a mobile app solution to dictate into the SRT.

The i-PARIHS Framework

Implementation science frameworks outline important considerations required for successful implementation of innovations15. While many theories, frameworks, and models have populated this space over time, these tools have not been fully utilized to guide the implementation of health IT and address the contemporary issues in health care16,17. To our knowledge, there have not been studies that have explored the role of implementation science frameworks for planning, developing, implementing, and evaluating interventions that aim to address EHR-related burden in the
mental health setting. Hence, there is an opportunity to leverage this source of knowledge to address issues related to documentation burden.

The integrated Promoting Action on Research Implementation in Health Services (i-PARIHS) framework aims to guide usage of evidence in implementing healthcare technologies. Building on the original PARIHS framework, this revised iteration focuses on achieving the project’s implementation goals through interactions that occur across the technology, environment, and the end-users. In particular, the framework outlines four essential components (facilitation, innovation, recipients, context) required for successful implementation (Table 1). The i-PARIHS framework has been used to support implementations of digital health technologies across primary care, acute care, and behavioural health.

Table 1. Overview of the components of the i-PARIHS framework.

<table>
<thead>
<tr>
<th>Construct</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Facilitation</td>
<td>Processes and activities that mitigate and adapt to the characteristics of the innovation, recipient and context</td>
</tr>
<tr>
<td>Innovation</td>
<td>Characteristics of the innovation as it relates to the implementation</td>
</tr>
<tr>
<td>Recipients</td>
<td>Characteristics of the actors and individuals that are part of the implementation</td>
</tr>
<tr>
<td>Context</td>
<td>Characteristics of the background/contextual factors that may influence the implementation</td>
</tr>
</tbody>
</table>

Objectives

In this case study, we applied the i-PARIHS framework to identify key considerations that are associated with implementing SRT in reducing documentation burden in mental health settings. In particular, we explore the following research question: What are the challenges and concerns that hinder the adoption of SRT as they pertain to each of the components of the i-PARIHS framework? While these insights were used to inform our implementation plan, we also aim to demonstrate the value of implementation frameworks in navigating complex issues such as documentation burden and identify any barriers in applying implementation frameworks, models and theories to these issues in a meaningful manner.

Methods

In order to apply the constructs of the i-PARIHS framework to identify key considerations for implementing SRT in mental health settings, a cross-sectional study design was used to gain insights into the needs, experiences, and perceptions towards SRT.

Settings and Participants

The study was conducted at a large academic and teaching mental health hospital located in Toronto, Ontario with a comprehensive EHR system. The hospital has obtained the highest level of adoption (Stage 7) on the HIMSS Electronic Medical Record Adoption Model (EMRAM). Prior to the roll-out of the improved version of SRT, some physicians had access to an earlier version of the SRT. Physicians without access to SRT either typed directly into the EHR system or used a back-end transcription service.

Physicians, residents, and clinical fellows were eligible to participate in this project if they held a full-time or part-time role at the organization as of January 2020 and completed documentation in the EHR system. Physicians who held a casual status were considered to not have sufficient experience with the EHR and were not eligible to participate in this study.

Data Collection

An online REDCap survey was developed using questions from existing literature, the SRT vendor, and our previous work to explore the innovation construct of the i-PARIHS. Participants were asked to indicate their previous experience with SRT at the organization and perceptions towards SRT on a 5-point Likert scale (Table 4). To build on the survey findings and explore the other constructs of the i-PARIHS and its relevant content (e.g., functionality, usability of the innovation), participants had the opportunity to provide additional thoughts, challenges, and concerns through open-ended responses and/or a follow-up interview. The survey was initially sent out in March 2020. Due to the COVID-19 pandemic, there was an organizational pause on all projects shortly after the first round of invitations. Data collection for the survey resumed in May 2020 and was completed in June 2020.
Follow-up interviews were conducted virtually by a member of the project team between April 2020 and September 2020 using a qualitative descriptive approach. The semi-structured interview guide (Table 2) queried their experience, perceptions, and concerns with using SRT. Like previous studies of this nature, interviews were not audio-recorded but notes were taken down by the interviewer. Participants of the follow-up interview were entered into a draw for three gift cards worth $25 CAD.

Table 2. Questions included in the semi-structured interview guide.

<table>
<thead>
<tr>
<th>Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. How do you prefer to document your patient notes?</td>
</tr>
<tr>
<td>2. What do you know about [SRT name] as a method of patient documentation?</td>
</tr>
<tr>
<td>3. What led you to [not] use [SRT name]?</td>
</tr>
<tr>
<td>4. How do you feel about using [SRT name] as a mode of documentation?</td>
</tr>
<tr>
<td>5. What do you look forward to about using [SRT name]?</td>
</tr>
<tr>
<td>6. What would be some of the barriers and challenges switching to [SRT name]?</td>
</tr>
<tr>
<td>7. If you were to switch to [SRT name], how can we better support you?</td>
</tr>
</tbody>
</table>

Data Analysis

The i-PARIHS framework guided our data analysis. Descriptive statistics (e.g., means, standard deviation) were used to characterize the sample and perceptions towards SRT. Differences in ratings amongst physicians with/without previous experience with SRT were determined using Kruskal-Wallis tests in SPSS Version 25. Deductive thematic analysis was conducted in NVivo 12 using the four constructs of the i-PARIHS framework where a single researcher coded the data. Investigator triangulation (including feedback from on-site experience) was used to enhance the trustworthiness of the qualitative findings. Since the survey was anonymous, there was no direct comparison between interview participants and their survey responses.

Ethical Considerations

Ethical approval was obtained through the quality improvement project review board at the Centre for Addiction and Mental Health. Numerous measures were taken to ensure the privacy and confidentiality of participants in this project. The survey responses were collected in an anonymous manner and the voluntary nature of participating in this project was reiterated in the email invitation. All survey and interview data is stored on the organizational secure server. Moreover, for the interviews, only data deemed relevant to the use and implementation of the SRT was coded and included in the data analysis. The interview data was also de-identified prior to the analysis.

Results

Seventy-eight individuals responded to the survey and 16 individuals completed the follow-up interview (Table 3). While respondents came from various clinical divisions, many respondents were staff physicians with outpatient duties. Less than half the respondents (n = 28) reported having previous experience with SRT. Findings from each i-PARIHS construct are discussed below.

Table 3. Demographics of respondents to the survey.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Number of Participants (N = 78) (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Role</strong></td>
<td></td>
</tr>
<tr>
<td>Clinical Fellow</td>
<td>2 (3%)</td>
</tr>
<tr>
<td>Resident</td>
<td>23 (29%)</td>
</tr>
<tr>
<td>Staff Physician</td>
<td>53 (68%)</td>
</tr>
<tr>
<td><strong>Primary Clinical Activities (multi-select)</strong></td>
<td></td>
</tr>
<tr>
<td>Emergency</td>
<td>22 (28%)</td>
</tr>
<tr>
<td>Inpatient</td>
<td>34 (44%)</td>
</tr>
<tr>
<td>Outpatient</td>
<td>60 (77%)</td>
</tr>
<tr>
<td>Outreach</td>
<td>5 (6%)</td>
</tr>
<tr>
<td>Telehealth</td>
<td>6 (8%)</td>
</tr>
<tr>
<td><strong>Clinical Division</strong></td>
<td></td>
</tr>
<tr>
<td>Addictions</td>
<td>5 (6%)</td>
</tr>
<tr>
<td>Adult Neurodevelopment and Geriatric Psychiatry</td>
<td>9 (12%)</td>
</tr>
<tr>
<td></td>
<td>Respondents without Experience (n = 50)</td>
</tr>
<tr>
<td>--------------------------</td>
<td>----------------------------------------</td>
</tr>
<tr>
<td>1. [SRT name] reduces my documentation time.</td>
<td></td>
</tr>
<tr>
<td>Strongly Disagree/Disagree</td>
<td>9 (18%)</td>
</tr>
<tr>
<td>Neutral</td>
<td>15 (30%)</td>
</tr>
<tr>
<td>Agree/Strongly Agree</td>
<td>24 (48%)</td>
</tr>
<tr>
<td>2. [SRT name] decreases my need to document after-hours.</td>
<td></td>
</tr>
<tr>
<td>Strongly Disagree/Disagree</td>
<td>10 (20%)</td>
</tr>
<tr>
<td>Neutral</td>
<td>14 (28%)</td>
</tr>
<tr>
<td>Agree/Strongly Agree</td>
<td>24 (48%)</td>
</tr>
<tr>
<td>3. [SRT name] uses more time than other methods of documentation.</td>
<td></td>
</tr>
<tr>
<td>Strongly Disagree/Disagree</td>
<td>23 (46%)</td>
</tr>
<tr>
<td>Neutral</td>
<td>13 (26%)</td>
</tr>
<tr>
<td>Agree/Strongly Agree</td>
<td>12 (24%)</td>
</tr>
<tr>
<td>4. [SRT name] is easy and intuitive to use.</td>
<td></td>
</tr>
<tr>
<td>Strongly Disagree/Disagree</td>
<td>10 (20%)</td>
</tr>
<tr>
<td>Neutral</td>
<td>16 (32%)</td>
</tr>
<tr>
<td>Agree/Strongly Agree</td>
<td>22 (44%)</td>
</tr>
<tr>
<td>5. I am interested in using [SRT name] for documentation.</td>
<td></td>
</tr>
<tr>
<td>Strongly Disagree/Disagree</td>
<td>8 (16%)</td>
</tr>
<tr>
<td>Neutral</td>
<td>8 (16%)</td>
</tr>
<tr>
<td>Agree/Strongly Agree</td>
<td>31 (62%)</td>
</tr>
</tbody>
</table>

The open-ended responses and follow-up interviews (Table 5) provided further insight into the perceptions observed in the survey. Many physicians without SRT experience frequently based their perceptions on testimonials voiced by other physicians. For example, in speaking about the accuracy of the dictations, a physician without experience (Quote 1A) was skeptical about the tool as they saw lukewarm success in their colleagues with using SRT. On the other hand, physicians with prior experience (Quote 1B) often cited their negative experience with previous versions of SRT, which generated reluctance towards using SRT. Thus, previous experience and testimonials from other physicians largely dictated their knowledge of SRT.

Some physicians also echoed specific challenges related to the features and functionalities of the SRT technology itself. One physician (Quote 2A) highlighted issues around the SRT recognizing homonyms and the added effort that is required to correct the mistakes. Alongside other issues such as dictating specific terminologies and formats (e.g., dates) (Quote 2B), these inconsistencies may lead to an unclear and poorly presented note for other clinicians.
Facilitation

In terms of facilitation, many respondents indicated that their lack of knowledge on the features of the SRT and how it should be used greatly hindered their decision to adopt SRT. For example, a physician (Quote 3A) suggested that addressing reluctance to change requires providing a tailored overview of the features of SRT, ideally in relation to how it may be used in practice. This may be accomplished through peer education where experienced users (e.g., super-users) of the system demonstrate how SRT can be used to perform and support the daily tasks of physicians (Quote 3B). As a result, ensuring physicians are aware of the features and value-add of adopting the SRT for their practice is critical to securing buy-in.

In terms of training and education, there were concerns around its content and delivery. Physicians had different preferred learning styles and constraints (e.g., time). For example, while some physicians preferred more independent and hands-off approaches for learning SRT (Quote 4A), others wanted more individualized training and support with mastering the skills of SRT (Quote 4B). There was also interest in a hybrid ‘curriculum’ approach that begins with self-learning of the basic features followed by training of more advanced features of the tool (Quote 4C). Several physicians (e.g., Quote 4D) also reiterated the value of ongoing technical and training support post-implementation. In summary, careful and continuous engagement with physicians is essential for obtaining the necessary buy-in for adoption and providing the required skills to use the solution.

Recipient

For the recipient construct, there was discussion on whether SRT is conducive to current practice workflows. Some physicians complete their documentation concurrently with the patient present and the additional layer of SRT can hinder the current workflow, and as one physician mentioned, may make the interaction uncomfortable for the patient (Quote 5A). Several additional considerations for implementing its usage with the organization’s EHR system were observed. For example, one physician highlighted how the setup of emergency department (ED) workflows at the organization would not be conducive to the use of SRT. In particular, the ED multidisciplinary assessment, which contains multiple tabs and many fields on each page, would hinder the use of SRT with more ‘clicks’ (Quote 6A). Other physicians also mentioned how SRT would conflict with their current habits and practices such as ‘copy-and-paste’ practices for documentation (Quote 6B). Thus, the individual practices, habits, and workflows of physicians and the environment can be a critical consideration in integrating SRT into practice.

Context

In terms of context, there was a focus on organizational factors and the setup of the environment. For example, many physicians complete their documentation at various locations, including at home, and some physicians question the equipment and setup needed to use SRT at each location (Quote 7A). Physicians also indicated that some of the current spaces are not set up for dictation of clinical notes. For example, from a privacy and confidentiality perspective, a physician (Quote 7B) highlighted the lack of quiet spaces to complete their documentation in between patients. This challenge was more prominent in certain environments, such as the emergency department.

Table 5. Sample quotes identified for each construct of the i-PARIHS framework  

<table>
<thead>
<tr>
<th>i-PARIHS Construct</th>
<th>Related Theme</th>
<th>Sample Quotes</th>
</tr>
</thead>
</table>
| Innovation         | Understanding of SRT is based on experience and testimonials | Quote 1A: “see my colleagues constantly editing the note manually when they dictate using [SRT name]” (Survey Respondent 69)  
Quote 1B: “Speech recognition software DOES NOT RECOGNIZE MY VOICE ever (including statements such as ’Hello, my name is...’), even with [SRT name].”  
(Survey Respondent 9) |
|                    | Specific challenges and concerns around the SRT | Quote 2A: “Certain words, homonyms (phrases that sound the same), the context of [SRT name] struggles to differentiate between the words. Constantly misidentify would have to put down mic, use mouse, edit the word, and back and forth. Not the same as human context in understanding.” (Interview Participant 4) |
| Facilitation                  | Understanding the features and value of SRT | **Quote 2B:** “... some challenge is retraining with your voice, issues, some glitches with current one, date format, patient names, numerics and abbreviations, becomes a pain and leave, reread the dictation and there’s some silliness.” (Interview Participant 11) |
|                             |                                              | **Quote 3A:** “Change is always difficult, used to doing workflow one way, so there would be some kind of learning curve in trying to incorporate. Find it hard to say without knowing how it functions, lots of hesitancy with not knowing what [SRT name] is.” (Interview Participant 12) |
|                             |                                              | **Quote 3B:** “It would be great if we can get a physician who is proficient with using [SRT name] to demonstrate the benefits and usage of [SRT name]. Some are concerned with [the] learning curve so it can be hard to sell. However, if we can see first-hand how it is used, that would be great.” (Interview Participant 6) |
| Education and Training      | Training requirements for adopting SRT      | **Quote 4A:** “Personally, I would like videos that they can go through at their own speed. Follow up with someone if there is troubles (helpline)” (Interview Participant 13) |
|                             |                                               | **Quote 4B:** “Make it really easy to switch, good to have tutorials and not just have online teaching, can sit down and work with someone on it [the training] either 1 on 1 or [in] small groups” (Interview Participant 12) |
|                             |                                               | **Quote 4C:** “The level of familiarity varies quite a bit with using technology and [SRT name]. It would be great to have self-learning because it is quite intuitive... Follow up then with advanced features.” (Interview Participant 6) |
|                             |                                               | **Quote 4D:** “I think I had someone spend 15-20 minutes with me when I first got it and when you know someone you can call asking for help and that was super helpful. Clear instructions on who to call and need to know exactly who to call.” (Interview Participant 16) |
| Recipient                   | Impact on physician’s workflow for delivering care | **Quote 5A:** “documenting in a room, kind of awkward while in room with patient, [it] would be weird” (Interview Participant 4) |
| Impact on physician’s       | Impact on physician’s workflow for documenting in the EHR | **Quote 6A:** “[SRT name] [is] not good for multidisciplinary assessment [with] so many tabs. It causes disruptions and lots of time to click than just dictation” (Interview Participant 6) |
|                             |                                               | **Quote 6B:** “For purpose of documentation, need to copy and paste a lot, a lot of this information has already been made available so copy and pasting is a lot faster and easier” (Interview Participant 2) |
| Context                     | Setup of required resources for SRT          | **Quote 7A:** “Communicating with IT around installing on different computers, microphone issues, shared offices, etc. I was unable to coordinate fixing the technical issues” (Survey Respondent 23) |
|                             |                                               | **Quote 7B:** “… if I have to dictate, then I would probably go see all my patients, [take] notes and then dictate at end of day. If on the run, [in a] nursing station with 5-6 team members, I probably wouldn’t document there” (Interview Participant 2) |

**Discussion**

As organizations explore ways to support the efficient and effective use of the EHR system, SRT continues to be an attractive option for many organizations. However, the lackluster adoption of an earlier version of SRT at our organization suggests the need for a thorough consideration of factors relevant for implementing SRT in practice. Implementation frameworks can offer great value in supporting the evidence-based and systematic implementation of innovation, yet their use has not been fully realized. In this study, we demonstrate the value of applying an implementation framework (i-PARIHS) in addressing the current challenges of documentation burden (Table 6).
Table 6. Summary of findings for each construct of the i-PARIHS framework 18.

<table>
<thead>
<tr>
<th>i-PARIHS Construct</th>
<th>Study Variables</th>
<th>Study Findings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Facilitation</td>
<td>• Required support (Interview Q7)</td>
<td>Customized, ongoing training and continuous physician engagement are necessary</td>
</tr>
<tr>
<td>Innovation</td>
<td>• Documentation preferences (Interview Q1)</td>
<td>Ensuring that the SRT technology recognizes multiple accents and is trained appropriately</td>
</tr>
<tr>
<td></td>
<td>• Knowledge of SRT (Interview Q2)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Perceptions of or experiences with SRT (Likert scale questions on survey, interview Q5)</td>
<td></td>
</tr>
<tr>
<td>Recipients</td>
<td>• Factors to using/not using SRT (Interview Q3, 4)</td>
<td>SRT may not be optimized for physicians’ workflows (i.e. those regularly completing multidisciplinary documentation, using copy-and-paste practices)</td>
</tr>
<tr>
<td>Context</td>
<td>• Barriers and challenges with using SRT (Open-ended responses on survey, interview Q6)</td>
<td>Ensuring that the SRT environment is appropriately set up to avoid privacy and confidentiality issues</td>
</tr>
</tbody>
</table>

Previous studies on using SRT have identified a catalogue of factors related to the adoption of SRT 12. However, these studies do not provide insight on the dynamic interactions that are likely required for successful implementation 30. The analysis of the collected data using the i-PARIHS framework 18 provides a unique perspective on understanding the relationships and issues related to the implementation of SRT. For example, in examining the facilitation and innovation constructs, there was a clear need for expectation-setting as part of communication 9, 30. While many of the respondents showed neutral perceptions towards its utility in reducing documentation time and ease of use, they remained interested in using SRT for documentation. Mastering SRT is associated with a steep learning curve, 30 however novel end-users may not be aware of this. This may lead to unrealistic expectations and early disengagement with SRT. Thus, there is a need to ensure that those in leadership positions have the appropriate knowledge and skills to manage expectations effectively 31.

Moreover, through the probing and questions of the interviews, the application of the framework highlighted the importance of implementing processes (facilitation aspects) that consider the challenges observed in the recipient, innovation, and context concepts. For example, in mental health, the suboptimal accuracy of dictating non-medical terminology (e.g., patient names) may translate to additional revisions or confusing documentation that can hinder patient safety 32. As a result, as part of the implementation process, providing adequate training on methods for correcting mistakes using SRT is instrumental to the ongoing improvement of the tool. 30 In addition, from a contextual perspective, the increased use of shared space can jeopardize the privacy of patients. Considering how SRT and similar technologies 33 can be used while maintaining patient privacy and confidentiality is needed 34. Lastly, from the perspective of physicians (the recipient), there are increasing concerns around how care can be delivered through and with digital technologies 35. As voiced by a few physicians, the use of SRT during the patient encounter can generate additional layers of complexities that hinder the therapeutic relationship with the patient 36. As such, identifying ways to maintain delivery of care while using SRT is needed to maximize its utility and value.

Implications for health care organizations

This study provides an example of how implementation frameworks can be applied in addressing EHR-related burden and clinician burnout and suggest that these frameworks can be valuable in implementation success. Healthcare leaders and administrators may consider leveraging a framework in guiding the implementation and evaluation of their own health informatics projects 17. While we selected the i-PARIHS 18 framework for this project, many other implementation frameworks could be used and the suitability of each alternative should be considered based on the objective and role of the project.

These findings also have implications for healthcare organizations interested in implementing SRT into their organization. From this study, it is evident that there are distinct nuances in psychiatry and mental health settings.
that may present a challenge for adopting SRT. For example, the use of unstructured, narrative documentation may necessitate robust training with the SRT to ensure that both medical and non-medical terminology are dictated correctly. Moreover, given the heightened privacy considerations around mental health issues, private offices or spaces to complete their documentation are needed.

Based on the findings from the application of the i-PARIHS framework, several recommendations are identified for implementing SRT in mental health settings. Foremost, given the additional training associated with accurate dictations by the SRT, there is a need to clearly communicate features and expectations associated with the ‘learning curve’ of adopting SRT. In particular, demonstrating how advanced features such as custom dictionaries and autotext can help address issues related to inaccuracies would be useful. This may be captured as part of a comprehensive training curriculum that includes both basic and advanced features of the solution. Dedicated workflows for optimizing the accuracy of dictations (e.g., the addition of commonly used non-medical terminology in the custom dictionary) may also be useful for supporting the ongoing optimization of SRT. Finally, identifying the workspace and workflow requirements to support the use of SRT is needed. This may include ensuring appropriate equipment and private spaces to complete documentation and implementing convenient locations to complete documentation throughout clinical settings.

Implications for Research

This study outlines opportunities for future research on the use of implementation frameworks for addressing EHR-related burden. With the numerous frameworks that are available, it may be useful to develop guidance on the selection and implementation of implementation frameworks on similar projects. As part of this work, building up a greater evidence base on the role of these frameworks for addressing unintended consequences arising from health IT use (e.g., administrative burden) is warranted.

In terms of the use of SRT in mental health settings, this study provides a baseline description of the perceptions and challenges towards adopting SRT for clinical documentation. Future studies may choose to explore how different implementation and optimization techniques can enhance the level of satisfaction and perceptions towards SRT for clinical documentation. The challenges identified from this study also provides the basis for building appropriate interventions to enhance implementation success. The impact of these interventions may then be evaluated using adoption metrics, EHR log data, and survey questions.

Limitations

Several limitations should be kept in mind when considering the results of this study. This study is only limited to the perceptions of physicians, residents, and clinical fellows from a single site. As the organization has already implemented the use of a comprehensive EHR system for many years, it is unclear if these perceptions are similar to those in organizations at a different stage of EHR adoption. More research is also needed to understand whether these perceptions can be translated to nurses and professional practice staff. Methodologically, this study was conducted using a cross-sectional design, and analysis was limited to a single implementation framework. The anonymous nature of the survey also prevented the ability to correlate the interview responses to the Likert scale and open-ended responses of the survey. In addition, the application of the implementation framework was only done as part of the analysis and not during the development and collection of the study. This may have hindered the value the i-PARIHS framework provided in addressing this issue. Lastly, the study was conducted when the COVID-19 pandemic was announced. Given the circumstances of the situation, we had a lower than expected response rate (~12% for staff physicians). With a smaller sample size, it is likely that there may be more participation from those with inherently greater interest for SRT (volunteer bias). Moreover, as the pandemic has led to dramatic shifts in the delivery of care, it has forced the uptake of various digital health tools including virtual care. It is unclear if the perceptions from this study have evolved since the rapid adoption of technologies during the pandemic.

Future Directions

As SRT continues to be an area of interest in addressing documentation burden, this study provides a foundation for further research on the use of SRT for clinical documentation in mental health settings. Foremost, it would be useful to explore the value of other implementation frameworks (e.g., RE-AIM) in supporting the mitigation of EHR-related burden. These frameworks may be considered in the implementation of other interventions for documentation burden such as EHR optimization sprints. In terms of implementing SRT in mental health settings,
there is a need to explore appropriate interventions to address the challenges observed in this case study. Building a strong body of evidence on SRT implementation in mental health settings can inform guidance on maximizing the value of these tools in reducing clinician documentation burden. Lastly, it may be useful to consider further sub-group analyses to characterize populations (e.g., demographics) that may be more/less reluctant to adopt SRT. Identifying these insights with a larger sample size for mental health as well as for other specialties (e.g., radiology) may be helpful for developing targeted and more effective interventions.

Conclusions

Through the application of the i-PARIHS framework on the implementation of SRT, this study demonstrates the value of applying well-established implementation frameworks in the implementation of digital health tools. Future implementations should consider the use of implementation frameworks to ensure success in using digital health to address challenging issues such as documentation burden and clinician burnout. In addition, several key considerations for implementing SRT in mental health settings were identified. Careful consideration of these challenges using robust education, support, and communication interventions are likely needed for project success.

Funding Sources

There were no formal sources of funding for this work, but in-kind contributions were provided by the Centre for Addiction and Mental Health.

Acknowledgements

The authors would like to thank all the physicians who kindly took the time to participate in this project. We would also like to acknowledge co-op students Dennis Hang, Hazel Walker, Kerry-Ann Smith and Surya Pandiaraju for their support with the initial stages of the project.

References


811
Understanding Heart Failure Patients EHR Clinical Features via SHAP Interpretation of Tree-Based Machine Learning Model Predictions

Shuyu Lu, MS¹, Ruoyu Chen, PhD¹,²,¹i, Wei Wei, PhD, Mia Belovsky, BS³, Xinghua Lu MD, PhD¹
¹Dept. Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, USA; ²Computer School, Beijing Information Science & Technology University, Beijing, China ³Sidney Kimmel Medical College, Thomas Jefferson University, Philadelphia, PA, USA

Abstract

Heart failure (HF) is a major cause of mortality. Accurately monitoring HF progress and adjusting therapies are critical for improving patient outcomes. An experienced cardiologist can make accurate HF stage diagnoses based on combination of symptoms, signs, and lab results from the electronic health records (EHR) of a patient, without directly measuring heart function. We examined whether machine learning models, more specifically the XGBoost model, can accurately predict patient stage based on EHR, and we further applied the SHapley Additive exPlanations (SHAP) framework to identify informative features and their interpretations. Our results indicate that based on structured data from EHR, our models could predict patients’ ejection fraction (EF) scores with moderate accuracy. SHAP analyses identified informative features and revealed potential clinical subtypes of HF. Our findings provide insights on how to design computing systems to accurately monitor disease progression of HF patients through continuously mining patients’ EHR data.

Introduction

Heart failure, also commonly referred to as congestive cardiac failure, is a clinical syndrome when the heart is unable to pump sufficiently to maintain blood flow to meet the body’s needs. According to CDC, HF is one of the most prevalent diseases and with highest mortality rate within the US, with the estimating that 6.2 million adults are affected. When a patient has a clinical encounter, their signs, symptoms, labs and other data is input into the EHR and the treatment plan is tailored to their specific physiologic condition. Thus, accurately monitoring disease states and adjusting the regimen in a timely manner can prevent or slow heart failure progression. The availability of an incoming stream of patient data including (but not limited to) weight and blood pressure readings, symptom monitoring and medication tracking has significantly lowered the mortality rates of heart failure patients, readmission rates, medical expenses, and improved patient satisfaction. Accurately monitoring disease progression and timely adjusting regimen to prevent or slow progression would improve HF patients’ life expectancy and quality. With prevalent availability of EHR, continuously mining of the EHR by computational agents to accurately detect and report disease progression will likely become a component of long term care of HF patients. In this study, we investigated the feasibility of using computational agents to assess disease stage and characteristic clinical features that provide information about disease and progression.

An objective measurement of heart function is the EF of left ventricle of heart. Broadly classified, there are 2 types of heart failure: heart failure with reduced ejection fraction (HFrEF) and heart failure with preserved ejection fraction (HFpEF). The difference lies in whether the systolic or diastolic capabilities of the left ventricle is affected, which is primarily measured by the left ventricle ejection fraction (LVEF or EF) score. The relation between EF score and HF is shown in Table 1. According to 2016 European Society of Cardiology (ESC) Clinical Practice Guideline, an EF score less than 40% typically suggests HFrEF, while an EF score greater than 50% is often a sign of HFpEF. As a distinct quantitative measure of heart failure stage, therefore, compared to other symptoms of heart failure, EF score is an effective

<table>
<thead>
<tr>
<th>EF Score</th>
<th>Pumping Ability of the Heart</th>
<th>Level of Heart Failure</th>
</tr>
</thead>
<tbody>
<tr>
<td>50%-70%</td>
<td>Normal</td>
<td>No HF/HFpEF</td>
</tr>
<tr>
<td>40%-50%</td>
<td>Slightly below Normal</td>
<td>Slight Symptoms of HF</td>
</tr>
<tr>
<td>35%-40%</td>
<td>Moderately below normal</td>
<td>Mild HFrEF</td>
</tr>
<tr>
<td>&lt;35%</td>
<td>Severely below normal</td>
<td>Severe HFrEF</td>
</tr>
</tbody>
</table>

Table 1: Relation between EF score and HF

¹This work is completed during Dr. Chen’s visit at the University Pittsburgh
variable to use when applying machine learning methods to predict a patient’s heart failure status. Currently, EF score is measured using an echocardiogram during inpatient/outpatient visits, but the frequency of echocardiogram can be few and far between in comparison to recordings of other clinical data. Furthermore, encounters of a HF patient with a health system are multi-faceted and occur in multiple settings, e.g., outpatient visits, nurse call, pharmacy visits, etc. All these encounters are usually recorded in the EHR of a comprehensive health system like the University of Pittsburgh Medical Center (UPMC). Therefore, EHR is an ideal resource for exploring indicators for diagnoses and outcomes, and we hypothesize that the continuous mining of EHR to detect disease progress would be of high clinical value in future.

Machine learning models have been increasingly applied to explore medical scenarios related to heart failure. In general, interpretable models are preferred over “black box” models in clinical settings. Among modern machine learning models, tree-based models, e.g., the XGBoost, is often favored in clinical settings, because they are easy to interpret, capable of handling missing values, and solve overfitting and underfitting problems effectively with the help of regularization methods. However, most earlier studies concentrate on model performance or feature importance, paying little attention to fully understanding and explaining the predictions with interpretable methods. That is, a human user not only is interested to know what features are informative with respect to a prediction task, but also would like to know how to interpret an observed value of a feature with respect to the prediction task. To this end, machine learning methods have been developed to identify informative features and their interpretations in a model-agnostic fashion. For example, a model-agnostic interpretation method, such as SHAP, can take a dataset and different prediction models as inputs, apply the models to the data, and subsequently discover the characteristics of data features in each prediction model (thus model-agnostic) different prediction models.

In this paper, we investigated the utility of XGBoost model in predicting heart failure stages, which is represented by EF scores based on structured EHR data. We evaluated informative features and investigated their interpretability and characteristics using the SHAP framework. Finally, based on the characteristics of features and their values, we applied unsupervised cluster learning to discover subtypes (subpopulations) among HF patients. To our best knowledge, few studies attempt to address the same questions as reported here, and we anticipate that our approach lays a foundation for future development of computational agents capable of monitoring HF patient disease progress by continuously mining the EHR data of health systems.

Materials and Methods

Data Collection and Preprocessing

All the EHR data were obtained from UPMC from 2014 to 2019 and contained clinical conditions of the patients who have been diagnosed with HF before, identified using ICD9/ICD10 codes (I428.* and I50.* respectively) which indicate heart failure. The original dataset consisted of 9 different CSV files that were extracted from the UPMC EHR system, including demographics (DEMO*), vitals (VL_*), labs (LB_*), medical dispenses (MD_*), medical fills (MF_*), medical orders (MO_*), order results (OR_*), problem lists (PL_*), and diagnoses (DL_*). Each patient had a unique ID, and each file was a collection of data from over 2,000,000 encounters from 60,835 unique patients. In order to build a profile for each patient, we cleaned the raw CSV files based on rules, extracted each patient’s record from the CSV files and finally aggregated the information. The workflow is illustrated in Figure 1:

The rules for data processing were:

1) For medical fills, medical orders, medical dispenses, problem lists and diagnoses, only keep the drug and disease names that appear over 10,000 times (10,000/2000,000 = 0.5%) in the dataset as valid features.

2) For numerical features like age, BMI, and blood pressure, in order to exclude outliers as much as possible, normalize the values and only keep the ones between 1% and 99% percentile, and values outside this range are then set to the value of 1 percentile (MIN) or 99 percentile (MAX).

3) The medical fills, medical orders, and medical dispenses are mixture of the National Drug Code (NDC) and the Anatomical Therapeutic Chemical (ATC) code. The problem lists and diagnoses are mixture of the ICD-9

---

[ii] Data acquired through the Health Record Research Request (R3) of University of Pittsburgh under the IRB# PRO18080460
and ICD-10 code. For the sake of consistency, all drugs NDC codes are mapped to ATC codes, and all ICD-9 codes are mapped to ICD-10 codes with the help of Apache Lucene.

After data cleaning, 1894 features from the 9 categories were retained and the values were stored tables. Details are available in Feature Table.

Figure 1: Data clean & aggregate workflow

The EF scores were directly obtained from patients’ echocardiogram reports. We extract the rest of structure EHR data of a patient within 45 days of EF measure, and we processed the EHR to derive a feature vector matching an EF measurement. Since each patient can have multiple EF measurement across time, and a patients disease may progress through such time, we treat EF measurements and match features from a patient that were separated by more than 180 days as ”independent” cases. Finally, we obtained 130,727 cases. We then split the dataset: 70% for training, 20% for validation and 10% for testing.

XGBoost Model for Regression

We applied XGBoost for our prediction task. XGBoost is an optimized distributed gradient boosting library and has been successfully applied in medical studies. Boosting algorithms is to combine weak classifiers together to form a powerful one. XGBoost is a boosting tree model that combines many CART regression tree models.

The regression tree model attempts to predict the original EF score with structured tabular EHR data. We tuned several hyperparameters with the evaluation of 5-folder cross validation for the XGBoost model before reaching the final model. The parameters are tuned with the sklearn.model_selection.GridSearchCV package, with coordinate descent as the strategy. The final tuned parameters are listed in Table 2:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>n_estimators (# of trees)</td>
<td>100</td>
</tr>
<tr>
<td>max_depth</td>
<td>3</td>
</tr>
<tr>
<td>eta (learning_rate)</td>
<td>0.35</td>
</tr>
<tr>
<td>min_child_weight</td>
<td>1</td>
</tr>
<tr>
<td>col_sample_by_tree</td>
<td>1</td>
</tr>
<tr>
<td>col_sample_by_level</td>
<td>1</td>
</tr>
<tr>
<td>subsample</td>
<td>0.85</td>
</tr>
<tr>
<td>reg_alpha (L1 regularization)</td>
<td>0</td>
</tr>
<tr>
<td>reg_lambda (L2 regularization)</td>
<td>0.5</td>
</tr>
<tr>
<td>gamma</td>
<td>0</td>
</tr>
<tr>
<td>num_boost_round (# of boosting iterations)</td>
<td>500</td>
</tr>
</tbody>
</table>

SHAP for Model-Agnostic Interpretation

SHAP stands for SHapley Additive exPlanations. It is a game theoretic approach to explain the output of a machine learning model. It connects optimal credit allocation with local explanations using the classic Shapley value from game theory and their related extensions. SHAP assigns a unique SHAP value to each feature in a sample for an EF prediction. The SHAP value represents the deviation from the average predicted value for each case prediction brought by each feature. For our work, we applied the SHAP model to firstly generate SHAP values for all our test dataset cases, and then illustrated the SHAP summary plot and SHAP scatter plot using a publicly available SHAP API for a global understanding of our dataset.

T-SNE clustering with SHAP values

T-SNE (t-distributed Stochastic Neighbor Embedding) is a visualization machine learning algorithm based on stochastic neighbor embedding. In particular, it models each high-dimensional object with a 2D or 3D point, with similar
cases placed as nearby points in t-SNE space.

For our work, after assigning a SHAP value for each feature of a subject, we represented a sample using a feature vector of SHAP values and clustered data points accordingly. We examined whether this approach could reveal different subtypes of HF patients by inspecting the clusters in comparison to clustering samples in original feature space. We applied t-SNE algorithm to map the 1894 features in 2D to visualize the clustering results.

Model implementing details

The experiments were conducted on an on-prem server of 72 of CPU(Intel(R) Xeon(R) Gold 6140 CPU @ 2.30GHz). The operating system was Ubuntu 18.04.4 LTS. All the codes for the task were written in Python 3.6.

The XGBoost model was implemented with xgboost 1.1.1 and scikit-learn 0.23; the SHAP interpretation was implemented with shap; t-SNE algorithm was implemented with scikit-learn.manifold.TSNE, where the perplexity was set to 100.

Results

Predicting performance of XGBoost model

The performance for XGBoost regression is $RMSE = 12.6303 \pm 0.00201$ (95% CI) with 100 random attempts on validation dataset. We evaluated the correlation between the predicted EF and real EF scores (Figure 2A), of which $R^2 = 0.264$, with a $p < 10^{-32}$. Figure 2B, shows the first tree (the most prominent out of 100 trees), which illustrates what features are utilized by the tree and how it splits samples to make final predictions. For example, the first feature for split is the DEMO_GENDER, which indicates that gender is a very important feature for predicting EF among HF patients. Note that our model only used structured data from the EHR to predict a numeric score with moderate accuracy. Besides, The predictions clearly follows correct trends.

Interpretation with XGBoost and SHAP

The XGBoost model and SHAP evaluate features from different perspectives. The XGBoost uses the coverage to reflect the importance of a feature, which denotes the percentage cases in which a feature is utilized during the decision pathway in making the final prediction. On the other hand, SHAP analysis assigns a SHAP value for a feature in each case (i.e., case-specific), which reflects the impact on the feature (measured as deviation from the mean predicted value) when different combinations of features (including or excluding the feature of interest) are used to predict the target value of a case. The important features provided by two methods are shown in Figure 3, where Figure 3A is the feature importance generated with XGBoost with coverage $\geq 0.01$, and Figure 3B is the top 20 most important features according to the SHAP analysis.

Interpretation of SHAP scores is as follows. For the SHAP value plot, each row presents the distribution of SHAP values assigned to a feature across all cases. The $x$-axis denotes the SHAP value, and the unit reflects how much the presence/absence of a feature with a particular value in a case will lead to deviation of predicted EF score (the unit is %) from the mean of prediction values using all possible combinations of feature sets from the case. The pseudo-color of a data point indicates the value of the feature of interest in a case. The further a point deviates from the mean of

---

https://github.com/Frank-LSY/XGB-SHAP-EHR-EF
predictions (which is 0), the more impact the features have on the prediction in the case. Therefore, a positive SHAP value that is on the right side of the mean on the x-axis, indicates the feature with the value in a case leads to a target value above the average predicted value, and below the average value on the left side. For example, for VL_BMI, the patient’s SHAP value is positive when the BMI is higher, and negative when the BMI is lower (except for extreme cases). That is, in the SHAP interpretation for our prediction model, heart failure patients tend to have EF scores higher than average if the BMI value is high and vice versa.

As shown in Figure 3C, the two methods identified a common set of informative features as well as some disjoint features, but the overall ranking of the features is similar. Firstly, the most important features that have highest coverage for XGBoost feature importance and highest SHAP value impact is DEMO_GENDER. If we dive deeper, like what is shown in Figure 4, we can see that female patients tend to have about a 5% higher EF score compared to male patients (female is represented as 0 and male as 1). Additionally, both XGBoost and SHAP interpretations treat DI255 (Ischemic Cardiomyopathy), DI428 (other Cardiomyopathies) and DI429 (Cardiomyopathy, unspecified) as critical diagnoses; MD_C03CA (Sulfonamides, plain) as critical medical dispenses, OR_MITRAL_REGURGITATION (Mitral valve regurgitation) as critical order results, if presented. According to Figure 3B, BP_SYSTOLIC (systolic blood pressure) and BP_DIASTOLIC (diastolic blood pressure) all have relatively high SHAP values, but their contributions to the prediction are quite the opposite. As shown in Figure 5, the model tend to assign higher EF score value predictions if patients have higher systolic blood pressures, while assign lower EF score value predictions if patients have higher diastolic blood pressures.

Sample clustering analysis with t-SNE and SHAP value

SHAP analysis provides a new perspective for inspecting data points: it shows the case-specific impact of each feature of a data point. This provides us with an opportunity to inspect whether different cases share a common pattern (joint...
distribution) of SHAP scores that characterize a subset of cases, which would provide a new perspective to identify patients that may share a common underlying disease mechanism. We applied the t-SNE algorithm to visualize the distribution of samples in the original feature space as well as in SHAP score space. For the former, each case was represented in the original feature space, and for the latter, each case’s SHAP values were used as input features, and t-SNE projected the data points from both representation into a 2D space.

Figure 6: Clustering HF patients in different feature spaces

Figure 6 shows the general clustering result with original feature values and SHAP values of all features respectively. The color represents EF score for each case. Compared with Figure 6A, which is the clustering of samples based on original feature values, clustering based on SHAP values (Figure 6B) makes the data points with similar EF values closer on the plot. More specifically, patients with high EFs (HFpEF) appear to be evenly distributed among the patients with reduced EFs (HFrEF) in the t-SNE results based on the original features, whereas HFpEF patients tend to be more clustered together when SHAP feature values were used as inputs. Furthermore, the HFpEF patients form sub-clusters in the SHAP-derived t-SNE space, which indicates that there are distinct combinations of SHAP values (thereby clinical features) that were detectable by the t-SNE algorithm. In summary, representing data points in the SHAP space revealed characteristics of samples that were not detectable in the original data space. We then set out to investigate which features contributed to the sub-clusters in the SHAP-derived t-SNE analysis. For the 10 graphs in this section in Figure 7, we plotted each data point in the same position as they were in Figure 6B, and we used pseudo-color to illustrate the original values of different features in each sub-plot.

When inspecting the combination of features of different sub-clusters, it is interesting to note that gender appears to a major factor that leads to the division of two large sub-populations (Figure 7A). This finding agrees with the finding from XGBoost analysis as mentioned previously: gender is the most important feature to be considered when the XGBoost model trying to predict the EF score. The separation of patients according gender indicates that patients with different genders tend to have distinct combinations of characteristics of other features, which suggests there is a major difference in disease mechanisms of HF patients of different genders.

It can also be seen that mitral regurgitation is present in many patients except few smaller clusters. The presence of mitral regurgitation is usually associated with dilated left ventricles, which is frequently associated with reduced EF (HFrEF). For the rest of sub-clusters, each plot has at least one unique feature that clearly assumes a value different from that of the majority of the patients, reflecting the importance of such a feature in differentiate sub-populations of HF patients. It also provides a more in-depth understanding of the particular patterns of feature combinations and how they may reveal distinct disease mechanisms underlying these subgroups.
In this study, we showed that the XGBoost regression model could be trained to predict EF score (an objective measurement of heart function) with fair performance, only based on structured EHR data. Furthermore, we showed that SHAP analysis revealed information and patterns that could not be easily acquired through analyzing the original features identified by XGBoost. The results are encouraging in that they demonstrated the feasibility of monitoring HF patient disease stage, thereby progression of the disease, of HF patients through mining the EHR data. We anticipate that with more training cases and more information from detailed clinical notes that provide the information of symptoms and signs of HF patients, the accuracy of the predictive model can be further improved and eventually make such a system clinically applicable. Given the dominant impact of HF on human mortality, a small improvement in monitoring patient disease progression can be translated into a significant improvement in overall patient outcomes.

Some of our findings are confirmed by previous studies, which indicate the validity of our approach in discovering patterns and important features. For example, the 5% difference in EF values between men and women was reported in several publications. According to Chung et al. who reviewed 1435 women and 1183 men cMRI, and concluded whether there is heart failure or not, females have a median EF score of 75% and males have a median of 70%, $p < 0.001$. This is consistent with the 5% gender difference in our XGBoost predicted results. Moreover, earlier studies also reported that in patients with heart failure of different genders, females are more likely to be diagnosed with HFpEF, whereas males are more likely to be diagnosed with HFrEF. Some other recent findings also demonstrate that there are significant differences in the incidence, prevalence, disease course and pathophysiology of heart failure between biological males and females. Females tend to survive longer after a heart failure diagnosis and more often have diastolic dysfunction, while males have a overall higher incidence of heart failure. Additionally, the macro and microvascular pathology underlying the etiology of the heart failure has been proven to vary between males and females as well.

Our study also highlighted trends that have not yet been confirmed in previous publications: the relation between blood pressure and EF score. According to Katsuya et al., among patients with acute heart failure syndromes (AHFS), it has been reported that those with a reduced left ventricular ejection fraction (LVEF) tend to be hypotensive or normotensive, whereas those with a preserved LVEF tend to be hypertensive. Their study which evaluated 4831 patients led to the conclusion that patients with an admission SBP $< 120$ mmHg were more likely to have a reduced LVEF than a preserved LVEF. In contrast, patients with an admission SBP $\geq 120$ mmHg were equally likely to have a preserved or reduced LVEF, indicating that there was no relation between a higher admission SBP and the LVEF. However, we did not find references that suggest the changes in systolic and diastolic blood pressure would have opposite effects on EF score. Such a combination of characteristics in these features are apparent in our analyses. The discovery of such combination patterns, thereby sub-populations of patients, not only prompts further investigation.

---

**Discussion**

**Findings**

In this study, we showed that the XGBoost regression model could be trained to predict EF score (an objective measurement of heart function) with fair performance, only based on structured EHR data. Furthermore, we showed that SHAP analysis revealed information and patterns that could not be easily acquired through analyzing the original features identified by XGBoost. The results are encouraging in that they demonstrated the feasibility of monitoring HF patient disease stage, thereby progression of the disease, of HF patients through mining the EHR data. We anticipate that with more training cases and more information from detailed clinical notes that provide the information of symptoms and signs of HF patients, the accuracy of the predictive model can be further improved and eventually make such a system clinically applicable. Given the dominant impact of HF on human mortality, a small improvement in monitoring patient disease progression can be translated into a significant improvement in overall patient outcomes.

Some of our findings are confirmed by previous studies, which indicate the validity of our approach in discovering patterns and important features. For example, the 5% difference in EF values between men and women was reported in several publications. According to Chung et al. who reviewed 1435 women and 1183 men cMRI, and concluded whether there is heart failure or not, females have a median EF score of 75% and males have a median of 70%, $p < 0.001$. This is consistent with the 5% gender difference in our XGBoost predicted results. Moreover, earlier studies also reported that in patients with heart failure of different genders, females are more likely to be diagnosed with HFpEF, whereas males are more likely to be diagnosed with HFrEF. Some other recent findings also demonstrate that there are significant differences in the incidence, prevalence, disease course and pathophysiology of heart failure between biological males and females. Females tend to survive longer after a heart failure diagnosis and more often have diastolic dysfunction, while males have a overall higher incidence of heart failure. Additionally, the macro and microvascular pathology underlying the etiology of the heart failure has been proven to vary between males and females as well.

Our study also highlighted trends that have not yet been confirmed in previous publications: the relation between blood pressure and EF score. According to Katsuya et al., among patients with acute heart failure syndromes (AHFS), it has been reported that those with a reduced left ventricular ejection fraction (LVEF) tend to be hypotensive or normotensive, whereas those with a preserved LVEF tend to be hypertensive. Their study which evaluated 4831 patients led to the conclusion that patients with an admission SBP $< 120$ mmHg were more likely to have a reduced LVEF than a preserved LVEF. In contrast, patients with an admission SBP $\geq 120$ mmHg were equally likely to have a preserved or reduced LVEF, indicating that there was no relation between a higher admission SBP and the LVEF. However, we did not find references that suggest the changes in systolic and diastolic blood pressure would have opposite effects on EF score. Such a combination of characteristics in these features are apparent in our analyses. The discovery of such combination patterns, thereby sub-populations of patients, not only prompts further investigation.
of potentially distinct underlying disease mechanisms but also suggests that tailored prediction/monitoring models should be developed for different sub-populations (a mixture of expert models), to enhance their performance.

**Limitations**

This is an early attempt at using ML models to detect the stage of HF in patients. Currently, the model only utilizes the structured data from the EHR, missing a significant amount of information from clinical notes. Besides, the features we selected according to our feature engineering approach are not necessarily matching exactly with heart failure. This is why our XGBoost regression model only achieves fair performance. One future direction should be extracting an informative representation of symptoms and signs associated with HF to enhance the accuracy of the predictions. Another limitation is that the current model does not attempt to model the temporal trajectory of heart function, which is important for targeting earlier interventions in order to improve clinical outcomes. This will be addressed in future studies.

**Conclusion**

With the XGBoost model, SHAP interpretation and unsupervised clustering visualization, we can 1) predicted EF score from tabular EHR data with decent performance; 2) generated interpretations for both the XGBoost model and dataset; and 3) classified the subgroups of HF. The generated interpretations are consistent with HF diagnosis guidelines and human intuition. This article provides a basic understanding of certain variable associations with heart failure. The model demonstrated the variables such as gender, blood pressure, age, pulse, BMI, some diagnoses (miscellaneous cardiomyopathies), and medications (sulfonamides, alpha and beta blocking agents) all have an impact on heart failure stage. To a large extent, this indicates that the future use of machine learning models to construct clinical decision aids related to heart failure is justifiable and feasible.

**Acknowledgements**

This work is partially support by NIH grant (R01LM012011) to LX. The authors would like to thank Mr. Zheng Li for discussions.

**Author Contributions**

Lu, S conceived the study under advice of Lu, X. Lu, S designed and implemented the programs, performed data analysis, and drafted the manuscript. Chen, R contributed to data processing. Wei, W participated in study design and literature review. All authors contributed to writing and editing of the manuscript.

**References**


27. Release Highlights for scikit-learn 0.23 — scikit-learn 0.23.2 documentation. (n.d.). Retrieved November 29, 2020, from https://scikit-learn.org/stable/auto_examples/release_highlights/plot_release_highlights_0.23_0.html
Identifying ARDS using the Hierarchical Attention Network with Sentence Objectives Framework

Kevin Lybarger, PhD, Linzee Mabrey, MD, Matthew Thau, MD, Pavan K. Bhatraju, MD, MSc, Mark Wurfel, MD, PhD, Meliha Yetisgen, PhD
University of Washington, Seattle, WA, USA

Abstract

Acute respiratory distress syndrome (ARDS) is a life-threatening condition that is often undiagnosed or diagnosed late. ARDS is especially prominent in those infected with COVID-19. We explore the automatic identification of ARDS indicators and confounding factors in free-text chest radiograph reports. We present a new annotated corpus of chest radiograph reports and introduce the Hierarchical Attention Network with Sentence Objectives (HANSO) text classification framework. HANSO utilizes fine-grained annotations to improve document classification performance. HANSO can extract ARDS-related information with high performance by leveraging relation annotations, even if the annotated spans are noisy. Using annotated chest radiograph images as a gold standard, HANSO identifies bilateral infiltrates, an indicator of ARDS, in chest radiograph reports with performance (0.87 F1) comparable to human annotations (0.84 F1). This algorithm could facilitate more efficient and expeditious identification of ARDS by clinicians and researchers and contribute to the development of new therapies to improve patient care.

Introduction

Coronavirus disease 2019 (COVID-19) is caused by infection with the severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) and is associated with high mortality.\textsuperscript{1} A high-risk complication of COVID-19 infection is the development of the acute respiratory distress syndrome (ARDS), which is characterized by severe inflammatory lung injury. Other common hospital diagnoses, such as sepsis, pneumonia, and trauma, are also associated with the development of ARDS. Interventions to prevent injury from invasive mechanical ventilation and differences in clinical management have improved clinical outcomes in patients with ARDS;\textsuperscript{2–4} however, ARDS is commonly under recognized by clinicians. In an epidemiologic study involving 500 intensive care units across 50 countries, over 40% of all ARDS cases were not recognized by clinicians, and the diagnosis of over 30% of ARDS cases was delayed.\textsuperscript{5} In another study, investigators demonstrated that delays in initiating evidence-based treatments was associated with increased hospital mortality in patients with ARDS.\textsuperscript{6}

The identification of ARDS requires the assessment of lung injury patterns in chest imaging. A primary contributor to undiagnosed ARDS is the challenge of incorporating radiologist-derived chest imaging findings into diagnostic algorithms for ARDS. Per the “Berlin Definition,” ARDS diagnosis requires:\textsuperscript{7}

- \textit{timing}: condition occurs within one week of a known clinical insult or new/worsening respiratory symptoms
- \textit{chest imaging}: bilateral opacities that are not fully explained by effusions, lobar or lung collapse, nodules, or masses
- \textit{non-cardiogenic edema}: alveolar infiltrates are not fully explained by cardiac failure or hydrostatic edema
- \textit{oxygenation}: oxygenation measurements meet defined thresholds (mild, moderate, and severe)

The \textit{oxygenation} component requires decreased oxygenation and is generally documented in structured data in the electronic health record (EHR). The \textit{non-cardiogenic edema} component requires an absence of hydrostatic edema, and the associated risk factors may be captured in structured admit diagnosis codes or the clinical narrative. The \textit{timing} component requires a proximal risk factor for respiratory failure, for example the presence of COVID-19, and may be documented through lab results or diagnosis codes. The information needed to assess the \textit{chest imaging} requirements is typically represented in chest radiographs (x-rays) and computed tomography images, as well as the associated free-text reports describing radiologists’ findings and interpretation. Data-driven computer vision approaches for directly analyzing the chest radiographs images are still in development and are computationally expensive. This work explores the automatic identification of the \textit{chest imaging} requirements for ARDS in free-text chest radiograph reports.

Natural language processing (NLP) explores the analysis, interpretation, and transformation of language. Information extraction is a common NLP task that automatically extracts structured representations from unstructured text. We use
information extraction techniques to identify descriptions of opacities (increased radiodensity), classify the opacities as parenchymal (indicative of alveolar edema/infiltrates) or extraparenchymal (outside the lungs or not indicative of alveolar edema/infiltrates), resolve sidedness (unilateral or bilateral), capture size information (small, moderate, or large), and indicate negation (“not present”). We developed detailed annotation guidelines that include summary document-level annotations and detailed relation annotations that characterize opacities. Using this novel annotation scheme, we created a new annotated corpus of 420 chest radiograph reports, referred to as the Pulmonologist Annotated Corpus (PAC). This work presents the Hierarchical Attention Network with Sentence Objectives (HANSO) framework, which is an end-to-end neural model that utilizes both the document-level and relation annotations. We introduce an approach for leveraging entity and relation annotations with noisy spans to improve document classification performance within the HANSO framework. We compare the performance of HANSO against two gold standards: manually annotated chest radiograph reports and manually annotated chest radiograph images. HANSO achieves very high performance in identifying the presence of bilateral infiltrates, a key indicator of ARDS, relative to both the annotated reports (0.87 F1) and annotated images (0.87 F1). HANSO also identifies factors that are less consistent for ARDS, specifically extraparenchymal opacities, with high performance (0.80 F1).

Related Work

Many works explore NLP information extraction techniques with radiology reports. Within this body of radiology research, several works explore the identification of pulmonary conditions in chest radiograph reports. Most prior pulmonory information extraction work implements discrete document classification models where labels are assigned at the document-level, without utilizing word-level annotations or predictions. Bejan, et al. identify pneumonia in chest radiograph reports using Support Vector Machines (SVM) with word n-grams, medical concepts, and other features. Yetisgen, et al. automatically identify acute lung injury in chest radiograph reports using Maximum Entropy (MaxEnt) models that utilize word n-grams and assertion predictions (present vs. absent). Afshar et al. and Mayampurath et al. predict ARDS in chest radiograph reports using word n-grams and medical concept features using discrete modeling approaches, including decision trees, k-nearest neighbors, naive bayes, logistic regression, and SVM. Mayampurath et al. achieves the best performance in predicting ARDS using unigram term frequency–inverse document frequency (TF-IDF) features with SVM, which we implement here as a baseline.

Some recent NLP work with chest radiograph reports utilizes continuous, neural modeling approaches. Datta et al. annotate approximately 2,000 chest radiograph reports using a detailed relation-based annotation scheme that characterizes radiology phenomena across multiple dimensions. Datta implements neural entity and relation extraction models, including a baseline model consisting of stacked bidirectional long short-term memory (bi-LSTM) and conditional random field layers, as well as transformer-based approaches using BERT and XLNet. Apostolova et al. investigate ARDS using both clinical text and structured EHR data from the MIMIC-III database, exploring ARDS likelihood, mortality, and risk factors using learned vector patient representations. Apostolova’s patient vectors incorporate information from clinical notes, diagnosis codes, and other structured data using Convolutional Neural Networks and Gradient Boosting Machine.

This work is differentiated from prior ARDS-related information extraction work in multiple ways. This work presents a new detailed annotation scheme that identifies indicators and confounding factors for ARDS, including document-level summary labels and detailed relation annotations describing the support or evidence for the document-level labels. It introduces a new end-to-end, neural multitask model that predicts the document-level labels and utilizes the detailed relation annotations to augment learning. Additionally, this work presents an approach for leveraging noisy entity and relation annotations.

Methods

Data

This work utilized two existing clinical data sets from the University of Washington Harborview and Montlake campuses. The first data set, Data set A, includes 831 chest radiograph reports for 173 patients from February-September 2020. Data set A includes patients that were being evaluated under suspicion for COVID-19 and admitted to a medical or trauma intensive care unit (ICU). Inclusion criteria were: (1) ICU admission; (2) suspicion for COVID-19; (3) inva-
sive mechanical ventilation; (4) presence of at least one partial pressure of arterial blood oxygen-to-fraction of inspired oxygen ratio (PaO2/FI02) less than 300 mmHg. For Data set A, an expert chest radiologist annotated 154 radiograph images using the Berlin criteria to identify patients with diffuse bilateral pulmonary opacities. The second data set, Data set B, includes 1,279 radiograph reports for 788 patients from March-November, 2020. Data set B includes all patients hospitalized with COVID-19, resulting in a broader patient population than Data set A with varying degrees of severity of illness from COVID-19 infection.

Annotation Scheme: We developed a detailed annotation scheme that facilitates the identification of lung infiltrates and extraparenchymal opacities. Table 1 summarizes the annotated phenomena, and Figure 1 presents annotation examples from the BRAT annotation tool. Each report was annotated with two categories of labels: document labels and relational labels. The document labels summarize the annotators’ overall assessment of each chest radiograph report with two multiclass labels: infiltrates – consistent with ARDS and extraparenchymal – less consistent with ARDS. The document classes are: none – insufficient information for assessment or absence explicitly stated; present – condition present but sidedness unknown; unilateral – explicitly one lung; and bilateral – explicitly both lungs. The string, “⟨⟨ INFILTRATES ⟩⟩ ⟨⟨ EXTRAPARENCHYMAL ⟩⟩,” was appended to each report to facilitate the assignment of these document labels.

The relation annotations are evidence for the document labels and include annotated spans and links between spans. Although the annotated spans are not necessarily noun phrases, we refer to the spans as “entities” here. The entity types include region, side, size, and negation. The annotation of region, side, and size includes an identified span and the assignment of a subtype label that normalizes the span contents, mapping the phrase to a clinically significant label. For example, all region entities include a subtype label of parenchymal or extraparenchymal. The negation entity only includes an annotated span without a subtype label, although the type label conveys the span meaning (i.e. “absent”). The relation annotations indicate whether a side, size, or negation entity are an attribute of a region entity. All attribute (attr) relation annotations are unidirectional, where the first entity in all relations has type region).

Table 1: Annotation guideline summary

<table>
<thead>
<tr>
<th>Category</th>
<th>Type</th>
<th>Subtypes</th>
<th>Span examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Document</td>
<td>infiltrates</td>
<td>none, present, unilateral, bilateral</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>extraparenchymal</td>
<td>none, present, unilateral, bilateral</td>
<td>–</td>
</tr>
<tr>
<td>Relational</td>
<td>region</td>
<td>parenchymal, extraparenchymal</td>
<td>“pulmonary infiltrates” or “lung disease”</td>
</tr>
<tr>
<td></td>
<td>side</td>
<td>unilateral, bilateral</td>
<td>“right” or “both sides”</td>
</tr>
<tr>
<td></td>
<td>size</td>
<td>small, moderate, large</td>
<td>“trace” or “small”</td>
</tr>
<tr>
<td></td>
<td>negation</td>
<td>–</td>
<td>“not present” or “no”</td>
</tr>
</tbody>
</table>

The annotation scheme provides the information necessary to categorize each chest radiograph report with respect to the radiologic criteria for ARDS; namely the presence of bilateral opacities that are not fully explained by effusions, lobar or lung collapse, nodules, or masses. The presence of opacities was qualified as infiltrates (indicative of alveolar process) and/or extraparenchymal (indicative of effusions, collapse, nodules/masses, and atelectasis) with additional annotation of any report text documenting laterality and size. This approach mirrors the clinical heuristic used by ex-
expert radiologists and pulmonary/critical care clinicians when assessing the likelihood that a chest radiograph indicates the presence of ARDS. In our annotation scheme the document labels, infiltrates and extraparenchymal, are the best indicators of ARDS, and the relation annotations are included to support the document labels.

Annotation Scoring and Evaluation: The goal of this work is to extract salient information from chest radiograph reports and convert it to a structured representation that will complement other types of structured clinical data (e.g. PaO3/FiO2 ratio) to predict ARDS. The assessment of annotator agreement and extraction performance focuses on the information in the annotation schema that is most relevant to the large-scale, automated assessment of ARDS. For each entity, the subtype label captures the important span information, such that the associated text span is less informative.

Figure 2 presents the same sentence annotated by two annotators. Both annotators label a region entity with subtype parenchymal that is connected to a side entity with subtype unilateral. Although the annotators label different spans for the region entities (“midlung, basilar opacities” vs. “opacities”), both annotations identify unilateral parenchymal opacities (i.e. opacities in one lung). For the purposes of predicting ARDS, these annotations are equivalent, even though there are span differences. For entities with equivalent type and subtype labels, the spans are evaluated under two criteria: any overlap and partial match. Under the any overlap criterion, spans are considered equivalent if there is at least one overlapping token, and the performance is assessed based on span counts. For the region annotation in Figure 2, the entity spans “midlung, basilar opacities” and “opacities” overlap, so there is one matching span. Under the partial match criterion, spans are compared at the token level to allow partial matches, and performance is assessed based on the number of matching tokens. For the region annotation in Figure 2, the entity spans have one matching token (“opacities”) and three mismatched tokens (“midlung, basilar”). There is only one relation type (attribute or attr), and two relations are equivalent if the entities paired by the attribute relation are equivalent under the any overlap criterion. Performance is evaluated using precision (P), recall (R), and F1-score (F1).

Annotation Statistics: The Pulmonologist Annotated Corpus (PAC) includes 420 chest radiograph reports, 120 from Data set A and 300 from Data set B. PAC was annotated by two pulmonary and critical care fellows, whom each annotated half the corpus. PAC has an 80%/20% train/test split. Figure 3 presents the histogram of the PAC document labels. The indication or reason for a chest radiograph in this patient population is a respiratory complaint, so positive labels (label ∈ {present, unilateral, bilateral}) are frequent: 75% of reports have a positive infiltrates label, and 43% have a positive extraparenchymal label. The corpus includes an average of 3.0 region, 1.3 side, 0.8 negation, and 0.2 size entities and 2.7 relations per report.

Annotator Agreement: As part of the annotation guideline development and annotator training, we doubly annotated 20 reports, assessed inter-annotator agreement, updated the annotation guidelines, and provided additional annotator training. At the conclusion of the project, we doubly annotated 10 additional reports, to assess the agreement for the annotated corpus. The agreement for these 10 reports is presented in Table 2. The agreement for both document labels is high (0.90 F1). The entity agreement under the any overlap criteria is also high (0.85-0.96 F1). The entity agreement using the partial match criteria remains high for negation and side entities; however it is relatively low for region. These results suggest the annotators are generally labeling the same phenomenon; however, they differ in the selected spans, similar to the example in Figure 2. Relation agreement is very high for region-negation entity pairs.
and lower for region-side pairs (0.76 F1). While the region span annotations are noisy, the entity and relation annotations, still contain useful information for assessing ARDS.

<table>
<thead>
<tr>
<th>Document label</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>extraparenchymal</td>
<td>0.90</td>
</tr>
<tr>
<td>infiltrates</td>
<td>0.90</td>
</tr>
</tbody>
</table>

(a) Document labels

<table>
<thead>
<tr>
<th>Entity</th>
<th>F1 any overlap</th>
<th>F1 partial match</th>
</tr>
</thead>
<tbody>
<tr>
<td>negation</td>
<td>0.96</td>
<td>0.92</td>
</tr>
<tr>
<td>region</td>
<td>0.85</td>
<td>0.49</td>
</tr>
<tr>
<td>side</td>
<td>0.89</td>
<td>0.84</td>
</tr>
<tr>
<td>size</td>
<td>*</td>
<td></td>
</tr>
</tbody>
</table>

(b) Entities

<table>
<thead>
<tr>
<th>Relation pair</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>region-negation</td>
<td>0.96</td>
</tr>
<tr>
<td>region-side</td>
<td>0.76</td>
</tr>
<tr>
<td>region-size</td>
<td>*</td>
</tr>
</tbody>
</table>

(c) Relations

Table 2: Annotator agreement. *no size spans were annotated.

Information Extraction

The likelihood of a patient satisfying the chest imaging requirements for the Berlin definition of ARDS can be estimated from the document labels, infiltrates and extraparenchymal. We introduce the Hierarchical Attention Network with Sentence Objectives (HANSO) framework in Figure 4, to predict these document labels and incorporate the relation annotation information. HANSO is a neural end-to-end, multi-task model that includes sentence encoding and document encoding layers. It builds on Yang’s hierarchical attention network (HAN). HAN aggregates word-level information to the sentence-level and then aggregates sentence-level information to the document-level. Sentences are encoded using a multi-layer network consisting of a recurrent neural network (RNN) and self-attention, and documents are encoded using separate RNN and self-attention layers operating on the encoded sentences. We build on HAN, incorporating sentence-level prediction tasks to augment the learning of the sentence representations. The sentence targets are derived from the relation annotations. The sentence and document encoding layers in Figure 4 are implemented for each of the document labels, infiltrates and extraparenchymal, with a shared input recurrent layer. HANSO omits the additional RNN included in the document encoding layer of HAN.

In our initial experimentation, we implemented a span-based relation extraction model for the entity and relation labels, similar to our previous work, and tried using the extracted relation information to improve the prediction of the document labels. However, the entity and relation extraction performance was insufficient to improve the prediction of the document labels, which are the most important labels in our annotation schema. Contributing factors to the low entity and relation extraction performance include the small data set size and the variability in the annotated spans, especially for region entities. The subtype labels in our annotation scheme (e.g. unilateral and bilateral for the side entities) normalize the span information, so the noisiness in the annotated spans does not negatively impact the informativeness of the annotations. However, this noise in the span annotation does negatively impact model learning and span prediction.

To utilize the detailed relation information, the relations are converted to a one-hot encoding for each sentence, capturing the salient relation and entity information without explicitly identifying entity spans. Figure 5 presents examples of this relations-to-sentence label mapping process. Relations consisting of region-side pairs are represented as the subtype label pairs: \{parenchymal-unilateral, parenchymal-bilateral, extraparenchymal-unilateral, extraparenchymal-bilateral\}. Relations consisting of region-negation pairs are represented as the region subtype label and negation type: \{parenchymal-negation, extraparenchymal-negation\}. While this approach does not explicitly capture span in-
formation, this sentence-level encoding of the relations creates a summary of the most important annotated phenomena within each sentence. The size entities are infrequent within the corpus and are omitted from experimentation.

Figure 5: Relation-to-sentence label mapping example

In the description of the HANSO framework below, the subscripts $i$, $j$, and $k$ indicate the $i^{th}$ BERT word piece position, $j^{th}$ sentence, and $k^{th}$ document. We only include the $i$, $j$, and $k$ subscripts below that are needed to resolve ambiguity.

Input encoding: Input documents are split into sentences and tokenized using spaCy. The default en_core_web_sm spaCy configuration is used, except that line breaks always indicate a sentence boundary. Sentences are mapped to contextualized word embeddings using Bio+Clinical BERT. The contextualized BERT word piece embeddings feed into a bi-LSTM without fine tuning BERT (no backpropagation to BERT). We tried several different architectures involving fine tuning BERT but these architectures did not out perform HANSO, which is likely due to the very small training set. The forward and backward states of the bi-LSTM are concatenated to form $h_i$ with size $1 \times l_h$, where $i$ is the word piece position.

Sentence encoding: Each sentence is represented as the attention-weighted sum of the word piece vectors. The bi-LSTM hidden state for each word piece is nonlinearly projected from size $l_h$ to $l_p$ as

$$u_i = \tanh(W_u h_i + b_u),$$

where $W_u$ is weight matrix and $b_u$ is a bias vector. Word-level attention weights, $\alpha_u$, are calculated using dot product attention as

$$\alpha_{u,i} = \frac{\exp(u_i^T z_u)}{\sum_i \exp(u_i^T z_u)},$$

where $u_i$ is the projected word piece input and $z_u$ is a learned vector of size $l_p$. The representation of sentence $j$ is calculated as

$$s_j = \sum_i \alpha_{u,i} h_i,$$

where $s_j$ has size $l_h$. A set of binary sentence-level prediction tasks, $R$, are incorporated based on the one-hot encoding of the relations described above. For each sentence-level task, $r \in R$, the sentence vector is nonlinearly projected from size $l_h$ to $l_p$ as

$$v_{r,j} = \tanh(W_v, s_j + b_{v,r}).$$

Label scores for task $r$ are calculated using a linear projection from size $l_p$ to 2 as

$$\psi_{r,j} = W_{\psi,r} v_{r,j} + b_{\psi,r},$$

where $\psi_{r,j}$ are the label scores for task $r$ and sentence $j$.

Document encoding: The same attention framework defined in Equations 1 and 2 is used to calculate the sentence-level attention weights, $\alpha_s$. Separate attention weights and bias vectors are learned for the sentence and document encoders. The sentence representation is nonlinearly projected from size $l_h$ to $l_p$ as

$$x_j = \tanh(W_x s_j + b_x).$$
Each document is represented as the attention weighted sum of the projected sentence vectors as
\[
d_k = \sum_j \alpha_{s,j} x_j,
\]
(7)
where \(\alpha_{s,j}\) is the attention weight for sentence \(j\) and \(d_k\) has size \(l_p\). The classes for the \textit{infiltrates} and \textit{extraparenchymal} labels are \{\textit{none, present, unilateral, bilateral}\}. Document label predictions are generated by linearly projecting the document vector from size \(l_p\) to \(l_d\) as
\[
\phi_k = W_d d_k + b_d
\]
(8)
where \(\phi_k\) are the label scores for document \(k\) and \(l_d\) is the document label set size.

Separate sentence and document encoders are implemented for the \textit{infiltrates} and \textit{extraparenchymal} document labels, utilizing a shared bi-LSTM input layer. For the prediction of the \textit{infiltrates} document label, the binary sentence prediction targets include: \textit{parenchymal-unilateral}, \textit{parenchymal-bilateral}, and \textit{parenchymal-negation}. For the prediction of the \textit{extraparenchymal} document label, the binary sentence prediction targets include: \textit{extraparenchymal-unilateral}, \textit{extraparenchymal-bilateral}, and \textit{extraparenchymal-negation}. To assess the contributions of the sentence-level objectives to document prediction performance, we implement HANSO without the sentence-level learning objective in Equation 5 ("HANSO lite") and the full HANSO model ("HANSO full").

**Baseline:** As a baseline for prediction performance, we implement a SVM model with unigram TF-IDF features, similar to Mayampurath et al.’s recent work predicting ARDS in chest radiograph reports.\(^\text{12}\)

**Experimental Setup:** Model hyperparameters are tuned through 3-fold cross validation on the PAC training set. The best performing models are applied to the withheld PAC test set. HANSO is implemented in PyTorch, and the SVM is implemented with scikit-learn.\(^\text{20,21}\) The HANSO configuration includes layer normalization at the bi-LSTM output and dropout (do) after the bi-LSTM and nonlinear projections in Equations 4 and 6. Additional hyperparameters include the number of epochs (ne), batch size (bs as document count), learning rate (0.002), maximum document length (35 sentences), maximum sentence length (30 tokens), hidden size (\(l_h = 100\)), and projection size (\(l_p = 100\)). The Adam optimizer is used, loss is summed across all targets, and the gradient norm is clipped (\(g_{\text{max}} = 1.0\)). For \textit{HANSO lite}, \(do = 0.1\), \(ne = 400\), and \(bs = 40\). For \textit{HANSO full}, \(do = 0.2\), \(ne = 150\), and \(bs = 10\). The SVM configuration includes the kernel ("rbf"), regularization (\(C = 10.0\)), and convergence tolerance (0.001).

**Results**

HANSO is trained on the PAC training partition and evaluated here using two approaches: \textit{text-versus-text} and \textit{text-versus-image}. In the \textit{text-versus-text} approach, the trained HANSO model is evaluated on the withheld PAC test set. The \textit{text-versus-text} approach is a typical NLP performance evaluation, where the model is trained and evaluated on annotated text. In the \textit{text-versus-image} approach, we apply HANSO to a set of chest radiograph reports for which the associated chest radiograph images are directly annotated by an expert radiologist. The \textit{text-versus-image} approach compares labels derived from text reports against gold standard image annotations.

**Text-versus-text**

In this subsection, we evaluate model performance on the withheld PAC test set. We present performance results for three models: SVM, \textit{HANSO lite}, and \textit{HANSO full}. To account for the variance associated with model random initialization, each model is trained on the training set 10 times and evaluated on the test set to generate a distribution of performance values. Table 3 presents the average performance across the 10 runs for the document labels, \textit{infiltrates} and \textit{extraparenchymal}. Performance is presented for each label and the micro average across labels ("micro"). \textit{HANSO full} achieves the best overall performance (F1 micro) for both \textit{infiltrates} and \textit{extraparenchymal} with significance (\(p < 0.05\)), demonstrating that the inclusion of the sentence objectives contributes to document prediction performance. \textit{HANSO full} achieves a statistically significant improvement in identifying \textit{bilateral infiltrates} relative to the SVM (\(p < 0.01\)), and the improvement of \textit{HANSO full} over \textit{HANSO lite} barely misses significance criteria (\(p = 0.07\)). Significance is assessed using a two-side t-test with unequal variance. The average performance across all sentence-level tasks in the \textit{HANSO full} runs is 0.83 F1. Considering the sentence-level tasks represent the summarization of the most salient relation information without including any span information, this performance is high.
Table 3: Document label prediction performance and gold standard label counts. † indicates the best performing model with significance ($p < 0.05$).

To assess the receiver operating characteristic (ROC) for HANSO’s ability to identify bilateral infiltrates, we train a separate HANSO model (HANSO full) with binary document targets ($0 = \text{not bilateral}$ and $1 = \text{bilateral}$), where the none, present, and unilateral labels are mapped to not bilateral. This binary variant achieves similar performance (0.88 F1) in identifying bilateral infiltrates as the multi-class models in Table 3. Figure 6 presents the ROC for bilateral infiltrates identification using the binary HANSO model. The ROC area under the curve (AUC) is 0.92. Optimizing the prediction threshold using Youden’s J statistic yields $J = 0.77$ at $FPR = 0.08$ and $TPR = 0.84$.

Text-versus-image

In this subsection, we compare the manual PAC annotations and automatically generated HANSO labels against annotated chest radiograph images. An expert radiologist annotated 154 chest radiograph images from Data set A with quadrant-level consolidation scores that can be mapped to the infiltrates labels none, unilateral, and bilateral, which we treat as the gold standard labels in this section. The annotated radiograph images correspond with 44 annotated reports in PAC (35 train and 9 test). The manual report labels are evaluated for the 44 reports in PAC that have a corresponding annotated images. The HANSO labels are evaluated for the 119 (154-35) annotated radiograph images not associated with reports in the PAC train set. Table 4 presents the performance of the manual and HANSO infiltrates labels. While the performance for none and unilateral is lower for both the manual and HANSO labels, the performance for bilateral is high for both the manual labels (0.84 F1) and HANSO labels (0.87 F1). ARDS diagnosis requires the presence of bilateral infiltrates, so the bilateral label performance is most important. HANSO achieves a sensitivity (recall) of 0.85 and specificity of 0.75 for bilateral in a one-versus-rest evaluation (bilateral vs. not bilateral). As the manual and HANSO performance is assessed using different samples, the significance of the performance differences cannot be assessed.
Conclusions

We introduce a new annotated corpus of chest radiograph reports, PAC, which includes document, entity, and relation annotations associated with ARDS. We also introduce the multi-task, end-to-end HANSO classification framework, which hierarchically encodes documents by encoding the word in sentences and the sentences in a document. Inter-annotator agreement for the PAC document labels is high (0.90 F1). The agreement for the entities indicates the annotators are generally identifying the same phenomenon in the chest radiograph reports (“any overlap” agreement 0.85-0.96 F1), although there is variability in the bounds of the annotated spans for the entities (“partial match” agreement 0.49-0.92 F1). The annotation scheme defines entities with type and subtype labels that normalize each span and capture the information most relevant to ARDS, so the variability in the bounds of the span annotations does not materially impact the clinical meaning of the annotations. However, this variability makes span extraction (entity recognition) challenging. To leverage the entity and relation annotations, we introduce an approach for mapping relations to a one-hot encoding of the entity pairs in each relation and use this one-hot encoding of the relations to create a set of sentence classification tasks. The one-hot encoding captures the most important annotated relation information, without requiring the prediction of entity spans. The primary objective of the HANSO framework is the prediction of the document labels associated with ARDS; however, HANSO includes a secondary objective associated with the prediction of the sentence-level one-hot encodings of the relations. The inclusion of the sentence-level objective increases performance, with significance, for the document labels; infiltrates increases from 0.71 to 0.79 F1, and extraparenchymal increases from 0.74 to 0.80 F1. The presence of bilateral infiltrates is predicted with very high performance (0.87 F1). HANSO predicts the one-hot encoded relations with high performance (0.83 F1), indicating the model is able to identify the key information from the fine grained annotations without explicit knowledge of span information. HANSO also outperforms a strong SVM baseline from recent ARDS information extraction work, with significance. We also assess the performance of manual (human) and automatic HANSO chest radiograph reports labels relative to annotated chest radiograph images. In the identification of bilateral infiltrates, HANSO achieves high performance against the annotated images (0.87 F1), which is comparable to the human performance (0.84 F1).

ARDS is a common complication of COVID-19 infection with high mortality; however, the identification of ARDS is often delayed or missed entirely. Delays in diagnosis lead to delays in evidence-based therapies that can improve clinical outcomes in COVID-associated ARDS. To our knowledge, this is the first study in COVID-19 that uses radiology reports to develop an automated NLP algorithm for identifying indicators of ARDS. The extracted ARDS indicators can be combined with additional clinical data to implement the real-time surveillance of ARDS in COVID-19 infected populations and ensure early implementation of evidence-based strategies for decreasing ARDS mortality. As next steps, we will complete an external validation of the data set and modeling framework to prepare for the deployment of an ARDS diagnostic tool. Specifically, the developed HANSO algorithm and trained model will be incorporated into a larger diagnostic tool that will be released and validated within the Electronic Medical Records and Genomics (eMERGE) Network to support pulmonary phenotyping efforts across different sites.

Acknowledgements

This work was supported by NIH/NHGRI (1U01 HG-008657), NIH/NLM Biomedical and Health Informatics Training Program (5T15LM007442-19), and NIDDK K23DK116967 (PKB). We want to acknowledge Sudhakar Pipavath, MD, for his contributions to the annotation of the chest radiograph images used in this study. Research and results reported in this publication was partially facilitated by the generous contribution of computational resources from the University of Washington Department of Radiology.

References


Toward Understanding Clinical Context of Medication Change Events in Clinical Narratives

Diwakar Mahajan\textsuperscript{1}, MS, Jennifer J. Liang\textsuperscript{1}, MD, Ching-Huei Tsou, PhD
IBM T.J. Watson Research Center, Yorktown Heights, NY

Abstract

Understanding medication events in clinical narratives is essential to achieving a complete picture of a patient’s medication history. While prior research has explored identification of medication changes in clinical notes, due to the longitudinal and narrative nature of clinical documentation, extraction of medication change alone without the necessary clinical context is insufficient for use in real-world applications, such as medication timeline generation and medication reconciliation. Here, we present a framework to capture multi-dimensional context of medication changes documented in clinical notes. We define specific contextual aspects pertinent to medication change events (i.e. Action, Negation, Temporality, Certainty, and Actor), describe the annotation process and challenges encountered while creating the dataset, and explore models based on state-of-the-art transformers to automate the task. The resulting dataset, Contextualized Medication Event Dataset (CMED), consisting of 9,013 medications annotated over 500 clinical notes, will be released to the community as a shared task in 2021-2022.

Introduction

An accurate medication history is foundational for providing quality medical care, allowing healthcare providers to better assess the appropriateness of current treatments, detect potential medication-related pathologies or symptoms, and direct future treatment options. Although medication information is captured in various sources of clinical data, in practice, providers often rely on structured medication orders due to easier access in electronic health record systems. However, many medication events are documented only in unstructured clinical notes\textsuperscript{1, 2}, which contain richer information but are difficult to search through especially at the point-of-care. In clinical practice, several scenarios may occur where a medication change is documented in clinical narratives but not in structured medication data. For example, if the patient already has a medication, the provider may instruct the patient to adjust the dosage or temporarily hold the medication without indicating the change in the medication order. Patient-initiated medication changes are also rarely captured in structured medication data, but would be documented in unstructured clinical narratives. Therefore, capturing medication event information from unstructured data within the patient medical record is required to achieve a full and complete picture of the patient’s medication history.

When extracting medication changes from clinical text, it is important to consider the surrounding contextual information because of the longitudinal and narrative nature of clinical documentation. Specifically, the longitudinal quality of clinical text results in documentation of events over the course of the patient’s medical history, from past and present events to future possible events. In addition, when documenting a clinical interaction with a patient, providers not only record what has occurred in the patient and the plan forward, but may also describe their clinical reasoning behind any medical decisions, including why certain treatment options were deferred. Such qualities of clinical documentation result in complex clinical events that would be insufficiently captured by extraction of medication change alone without consideration of the surrounding clinical context. This is especially true when developing a medication change extraction system to support real-world applications, such as medication timeline generation\textsuperscript{3, 4} or medication reconciliation\textsuperscript{5, 6}. For example, to generate a medication timeline, a system must extract not only the dosage adjustment action, but also place it in the correct point in time, i.e. knowledge of when the action is occurring. Similarly, medication reconciliation requires awareness of not just what medications were prescribed by healthcare providers, but also whether the patient is taking (or not taking) a medication, i.e. knowledge of patient-initiated actions.

To address the need for consideration of contextual information during medication change extraction, here we propose a conceptual framework to organize multi-dimensional context for medication events in clinical narratives, and present the resulting dataset – Contextualized Medication Event Dataset (CMED). CMED captures pertinent context for medication change events along five orthogonal dimensions (i.e. Action, Negation, Temporality, Certainty, and Actor), and

\textsuperscript{1}Equal contribution
will be released to the research community as a shared task in 2021-2022. To the best of our knowledge, CMED will be the first dataset on medication change events made available to the research community. The major contributions of this paper are:

- a dataset of 9,013 medication mentions annotated with contextualized medication change events over 500 clinical notes,
- insights into the language used by clinicians when documenting medication changes and the challenges in interpreting this language during the annotation process, and
- experimental results showing that support vector machines (SVM) and transformer-based models can partially automate this task, with error analysis providing insights into the most challenging aspects of this task.

Related Work

Previous attempts in classifying medication change events in clinical notes were driven by a variety of use cases. For example, some prior works focused only on specific medications, while others focused on specific types of medication changes but for all medications. Past works targeting specific medications include work on warfarin (labels on or stop)\(^7\), heart failure medications (labels active, discontinued, or negative)\(^9\), beta blockers (labels active, medication list, negated, discontinued, or other)\(^9\), and dietary supplements (labels continuing, discontinued, started, unclassified)\(^10\). In other works targeting medication changes but for all medications, Liu et al. focused only on medication discontinuation events\(^12\), Sohn et al. considered a greater range of medication changes (labels start, stop, increase, decrease, no-change)\(^13\), Lerner et al. also included labels that indicate sequential changes (labels start, start+stop, stop, continue, switch, decrease, increase)\(^14\), and Pakhomov et al. introduced temporal information in one of their labels (labels past, continuing, stop, start, not classified)\(^15\). Note that these works, driven by a variety of use cases, have resulted in a mixed set of labels, some of which may not cover all types of changes (e.g. labels used by Pakhomov et al. - past, continuing, stop, start, not classified\(^15\), do not differentiate between increase and decrease), or do not cover all aspects of a medication event (e.g. labels used by Sohn et al. - start, stop, increase, decrease, no-change\(^13\), cover all types of changes but do not provide temporal information).

In another body of work, attempts have been made to recognize context or assertions for medical concepts such as problems and tests\(^16\)–\(^19\). There have also been attempts at identifying negated medical concepts in clinical text\(^17, 20–23\). Although some of these works identify aspects such as certainty and negation in medical events, none of them have been applied to medication change events. Further, no previous work has attempted to identify the actor behind an event, which is especially important for medication change events due to implications for patient adherence. Thus, there is need for a more organized schema of label definitions that better contextualizes medication events.

Our work differs from previous research in several ways. First, we capture relevant contextual information on top of just identifying medication change in clinical narratives. Second, the contextual information captured is organized along multiple orthogonal dimensions. Finally, we will be releasing our annotated dataset on contextualized medication change events as part of a shared task.

Methods

Data and Annotation

For this study, we used the 2014 i2b2/ UTHealth Natural Language Processing shared task corpus\(^24–26\), which contain a total of 1,304 clinical notes over 296 patients. This corpus was selected due to its longitudinal nature, with 2-5 notes per patient, to allow potential future work in reconciling medication events extracted across different notes. To benefit from the longitudinal nature of this corpus while ensuring sufficient variation in our dataset, we first selected 44 patients from which all notes were chosen (total 199 notes), and then randomly selected 1-2 notes from each of the remaining 252 patients (total 301 notes). The resulting corpus consisted of 500 notes over 296 patients, of which 120 notes were doubly-annotated and adjudicated to measure inter-annotator agreement (IAA). Notes were annotated by a team of three annotators led by a physician. To assist the annotators, medications in the notes were pre-annotated using a medication extraction model\(^27\), and corrected if necessary by the annotators during the annotation process (i.e. incorrect pre-annotations were removed and missed medication mentions were added).
We define a medication change event as any discussion about a medication change for a given patient. The annotation process is as follows. For each medication mention, the annotator first determines whether a medication change is being discussed, and assigns it one of the following medication change event labels:

- **NoDisposition**: no medication change is being discussed, e.g. “pt is on Coumadin”, “continue lisinopril”
- **Disposition**: presence of a medication change being discussed, e.g. “Start Plavix”
- **Undetermined**: unclear if a medication change is being discussed and additional information is required to make the determination, e.g. “Plan: Lasix” – unclear if just stating a medication patient is on (NoDisposition) or starting a new medication (Disposition)

Next, for identified Disposition events, the annotator labels the clinical context for the event along five dimensions:

- **Action**: What is the change discussed? (Start, Stop, Increase, Decrease, OtherChange, UniqueDose, Unknown)
- **Negation**: Is the change being discussed negated? (Negated, NotNegated)
- **Temporality**: When is this change intended to occur? (Past, Present, Future, Unknown)
- **Certainty**: How likely is this change to have occurred / will occur? (Certain, Hypothetical, Conditional, Unknown)
- **Actor**: Who initiated the change? (Physician, Patient, Unknown)

Figure 1 summarizes the annotation process.

Some sample annotations are shown in Table 1. Note that a single medication mention can have 0, 1, or more associated Disposition events. For example, “May try amlodipine if develops cough” has 1 associated Disposition event for amlodipine (Action|Negation|Temporality|Certainty|Actor = Start|NotNegated|Future|Conditional|Physician), while “Was started on metformin at last visit but pt stopped after one week because of GI upset” has 2 associated Disposition events for metformin (Start|NotNegated|Past|Certain|Physician and Stop|NotNegated|Past|Certain|Patient). These dimensions were designed to extract the relevant clinical context for medication change events, allowing us to capture a wide variety of attributes, such as references to past events (Temporality: Past), treatments being considered but not yet decided upon (Certainty: Hypothetical), or episodes of patient nonadherence (Actor: Patient).

While defining this task, we identified certain nuances of the annotation process that led to specific guidelines in our annotation protocol, specifically, the effect of (1) surrounding contextual information and (2) external medication-specific knowledge on label assignment. For contextual information, because providers can copy-and-paste text from previous notes, a sentence that appears to be Present based on tense (e.g. “Start prednisone taper”), when found in the Past Medical History (PMH) note section, actually indicates a Past action. To balance the need for consideration of
Table 1: Sample annotations demonstrating how labels change depending on the surrounding text.

<table>
<thead>
<tr>
<th>Text</th>
<th>Event</th>
<th>Action</th>
<th>Negation</th>
<th>Temporality</th>
<th>Certainty</th>
<th>Actor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pt currently on <em>lisinopril</em></td>
<td>NoDisp</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Plan: incr <em>losartan</em> from 1 tab qd to bid.</td>
<td>Disp</td>
<td>Increase</td>
<td>NotNegated</td>
<td>Present</td>
<td>Certain</td>
<td>Physician</td>
</tr>
<tr>
<td>Patient also given a prescription for <em>Hctz</em> 12.5mg QD but told not to start it yet until his next BP check and start only if BP &gt; 140/85</td>
<td>Disp</td>
<td>Start</td>
<td>NotNegated</td>
<td>Future</td>
<td>Conditional</td>
<td>Physician</td>
</tr>
<tr>
<td>In ED, given <em>ativan</em> 1 mg IV x1</td>
<td>Disp</td>
<td>UnqDose</td>
<td>NotNegated</td>
<td>Past</td>
<td>Certain</td>
<td>Physician</td>
</tr>
<tr>
<td>She was experiencing a bad episode of dry cough so stopped taking <em>lisinopril</em></td>
<td>Disp</td>
<td>Stop</td>
<td>NotNegated</td>
<td>Past</td>
<td>Certain</td>
<td>Patient</td>
</tr>
<tr>
<td>Will hold off on empirically starting <em>abx</em> based on urinalysis</td>
<td>Disp</td>
<td>Start</td>
<td>Negated</td>
<td>Present</td>
<td>Certain</td>
<td>Physician</td>
</tr>
<tr>
<td>Pt’s <em>lasix</em> dose was increased from 80 mg to 120 mg for four days, then reduced back to 80</td>
<td>Disp</td>
<td>Increase</td>
<td>NotNegated</td>
<td>Past</td>
<td>Certain</td>
<td>Physician</td>
</tr>
</tbody>
</table>

contextual information while avoiding annotator error due to cognitive overload from having to remember the entire note, we asked annotators to consider only local context when determining the appropriate label, roughly defined as the immediate sentence containing the medication mention +/- 1 sentence and the note section. For knowledge of a specific medication’s attributes (e.g. drug class, route of administration) affecting label assignment, because some medications are only available by prescription, while others are available over-the-counter, knowledge of a specific medication’s availability can affect whether the Actor is more likely to be Physician or Patient. Similarly, some medications, such as Z-Pak, are dispensed as a set dosage pack with time-limited administration. Therefore, with that knowledge, the medication in a phrase such as “Plan: *Z-Pak*” should be annotated as Start|NotNegated|Present|Certain|Physician and Stop|NotNegated|Future|Certain|Physician. To best reflect the reality of the event, annotators were asked to annotate based on what they believe is going on, even if the assigned label depends on medical knowledge not in the text.

**System Description**

To automate the task of medication change classification, we split the task into five classification subtasks organized in a two-step process:

1. One subtask to classify each given medication mention into Disposition, NoDisposition, or Undetermined
2. Four subtasks to classify each given Disposition medication along four context dimensions: Action, Temporality, Certainty, and Actor. Negation was excluded from our experiments due to limited number of Negated instances.

For each subtask, we train a classification model based on two approaches: (1) SVM, and (2) transformer-based language models. As CMED will be used in a shared task in 2021-2022, we set aside 100 notes from our annotated data as blind data for the task. The remaining annotations in 400 notes are split into 75% for training, 5% for development and 20% for test. In addition to Negation Dimension, OtherChange for Action dimension and Unknown for all dimensions were excluded in our experiments due to the limited number of instances (<40).

**Support Vector Machines.** To explore the dataset and various dimensions, we elected to use a feature-based approach with a discriminative classification algorithm in our experiments. Our features were divided into the following classes: lexico-syntactic, window, dep-parse, note-section, and RxNorm features.

- **Lexico-syntactic.** Standard lexical features were utilized such as n-grams of whole, stemmed and lemmatized words, and part-of-speech tags.
- **Window.** We incorporated windowed lexico-syntactic features where we combined the positional information of word tokens with their lexico-syntactic features for the tokens that appear within of a window of 5 from the target medication mention in either direction. The window size of 5 was empirically determined.
• **Dep-parse.** We utilized features based on the dependency parse structure of the sentence containing the target medication, e.g. neighbors, parent and children of the medication concept in the dependency parse tree.

• **Note-section.** We experimented with a rule-based note section classification model that classifies each sentence based on the section headers and location of the sentence within the note.

• **RxNorm.** We leveraged features derived from RxNorm for the target medication such as Anatomical Therapeutic Chemical class, ingredient etc.

Experiments were performed using SVM with a linear kernel and one-vs-all classification strategy as implemented in scikit-learn. As initial experiments with different kernels (linear, polynomial and radial basis function) showed comparable performance, we chose linear kernel for all of our experiments. We used scispaCy for sentence segmentation and extraction of lexico-syntactic and dep-parse features. RxNorm features were extracted using the RxNorm public API. To deal with skew in class distribution, we adjusted the class weights to be inversely proportional to class frequencies.

**Transformer-based Language Models.** We experimented with state-of-the-art Bidirectional Encoder Representations from Transformer (BERT)-based language models, with the models pretrained on general-domain (BERT-base) and in-domain (ClinicalBERT) datasets. We formulated each of the five tasks as a sentence classification task by providing the surrounding sentence of the annotated medication as the context. In this process, we employed the pre-trained transformer to obtain a distributed representation, 0.2 dropout, and a fully connected layer of size 5 with softmax activation to make the classification for each of the five subtasks. We used the transformers package to tune our models with the train and development splits, and present our results on the test split.

We evaluated our two system approaches with the following experiments on the 20% test data, (1) SVM, (2) BERT-base and (3) ClinicalBERT. We present the micro and macro precision, recall, and F1 scores in the Results section. We also present a majority baseline on the training dataset with majority label for each dimension being chosen as the prediction for comparison against our model.

**Results**

*Inter-annotator agreement and dataset statistics*

Overall, our annotators achieved high IAA based on Cohen’s kappa. On determining whether a medication change was being discussed (Disposition vs NoDisposition vs Undetermined), IAA was 0.88 on 2,495 annotated medication mentions. For the 367 instances of agreed Disposition events, Action, Negation and Temporality had the highest IAA at 0.87, 0.83, and 0.94, respectively, while Certainty and Actor had lower IAA at 0.75 and 0.72, respectively.

The resulting dataset, CMED, consists of 9,013 annotated medication mentions over 500 clinical notes. As CMED will be used in a shared task in 2021-2022, we only report detailed statistics on the training dataset in this paper. The training dataset consists of 7,230 annotated medication mentions over 400 notes, with distribution for specific labels presented in Table 2.

<table>
<thead>
<tr>
<th>Task</th>
<th>Label</th>
<th>Count</th>
<th>Task</th>
<th>Label</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Event</td>
<td>NoDisposition</td>
<td>5260</td>
<td>Temporality</td>
<td>Past</td>
<td>745</td>
</tr>
<tr>
<td></td>
<td>Disposition</td>
<td>1413</td>
<td></td>
<td>Present</td>
<td>494</td>
</tr>
<tr>
<td></td>
<td>Undetermined</td>
<td>557</td>
<td></td>
<td>Future</td>
<td>145</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Unknown</td>
<td>29</td>
</tr>
<tr>
<td>Action</td>
<td>Start</td>
<td>568</td>
<td>Certainty</td>
<td>Certain</td>
<td>1177</td>
</tr>
<tr>
<td></td>
<td>Stop</td>
<td>341</td>
<td></td>
<td>Hypothetical</td>
<td>134</td>
</tr>
<tr>
<td></td>
<td>Increase</td>
<td>129</td>
<td></td>
<td>Conditional</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>Decrease</td>
<td>54</td>
<td></td>
<td>Unknown</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>UniqueDose</td>
<td>285</td>
<td></td>
<td>Unknown</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>OtherChange</td>
<td>1</td>
<td></td>
<td>Unknown</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Unknown</td>
<td>35</td>
<td></td>
<td>Unknown</td>
<td>2</td>
</tr>
<tr>
<td>Negation</td>
<td>Negated</td>
<td>32</td>
<td>Actor</td>
<td>Physician</td>
<td>1278</td>
</tr>
<tr>
<td></td>
<td>NotNegated</td>
<td>1381</td>
<td></td>
<td>Patient</td>
<td>107</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Unknown</td>
<td>28</td>
</tr>
</tbody>
</table>

Table 2: Label distribution for the training dataset
As observed in Table 2, less than 20% of medication mentions have an associated Disposition event. Further, a substantial percentage (7.7%) of medication mentions could not be resolved into Disposition or NoDisposition events due to the lack of sufficient information. Within the Action dimension, Start and Stop account for over 64.3% of Disposition events, UniqueDose for 20.2%, and titration events (Increase and Decrease) for 13%. Labels in the Temporality dimension reveal that over half of Disposition events occur in the Past (52.7%), which is reflective of the longitudinal and narrative nature of clinical notes.

Labels in the Temporality dimension reveal that over half of Disposition events occur in the Past (52.7%), which is reflective of the longitudinal and narrative nature of clinical notes. For Certainty, 16.6% of Disposition events are discussed in a Hypothetical or Conditional context. The prevalence of Past events as well as Hypothetical and Conditional events further confirms the need for contextualized medication event information. For Actor, as expected, the majority of medication changes in clinical text are initiated by healthcare providers (90.4%). Finally, Unknown in all dimensions, Negated in Negation dimension, and OtherChange in Action dimension are rare labels, each with less than 40 instances in the dataset.

### System Results

Table 3 shows the results of (1) majority baseline, (2) SVM, (3) BERT-base, and (4) ClinicalBERT against the test dataset. While SVM provides a reasonable baseline, ClinicalBERT yields the best results across all five subtasks.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Task</th>
<th>Event</th>
<th>P</th>
<th>R</th>
<th>F1</th>
<th>Action</th>
<th>P</th>
<th>R</th>
<th>F1</th>
<th>Temporality</th>
<th>P</th>
<th>R</th>
<th>F1</th>
<th>Certainty</th>
<th>P</th>
<th>R</th>
<th>F1</th>
<th>Actor</th>
<th>P</th>
<th>R</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>majority baseline</td>
<td>micro</td>
<td>0.73</td>
<td>0.73</td>
<td>0.73</td>
<td>0.41</td>
<td>0.41</td>
<td>0.41</td>
<td>0.54</td>
<td>0.54</td>
<td>0.54</td>
<td>0.83</td>
<td>0.83</td>
<td>0.83</td>
<td>0.92</td>
<td>0.92</td>
<td>0.92</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>macro</td>
<td>0.24</td>
<td>0.33</td>
<td>0.28</td>
<td>0.08</td>
<td>0.20</td>
<td>0.12</td>
<td>0.18</td>
<td>0.33</td>
<td>0.23</td>
<td>0.28</td>
<td>0.33</td>
<td>0.30</td>
<td>0.46</td>
<td>0.50</td>
<td>0.48</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SVM</td>
<td>micro</td>
<td>0.79</td>
<td>0.79</td>
<td>0.79</td>
<td>0.59</td>
<td>0.59</td>
<td>0.59</td>
<td>0.71</td>
<td>0.71</td>
<td>0.71</td>
<td>0.83</td>
<td>0.83</td>
<td>0.83</td>
<td>0.88</td>
<td>0.88</td>
<td>0.88</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>macro</td>
<td>0.62</td>
<td>0.63</td>
<td>0.63</td>
<td>0.50</td>
<td>0.51</td>
<td>0.50</td>
<td>0.60</td>
<td>0.59</td>
<td>0.59</td>
<td>0.59</td>
<td>0.53</td>
<td>0.56</td>
<td>0.63</td>
<td>0.68</td>
<td>0.65</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BERT-base</td>
<td>micro</td>
<td>0.88</td>
<td>0.88</td>
<td>0.88</td>
<td>0.75</td>
<td>0.75</td>
<td>0.75</td>
<td>0.81</td>
<td>0.81</td>
<td>0.81</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.92</td>
<td>0.92</td>
<td>0.92</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>macro</td>
<td>0.79</td>
<td>0.77</td>
<td>0.78</td>
<td>0.75</td>
<td>0.62</td>
<td>0.64</td>
<td>0.77</td>
<td>0.71</td>
<td>0.73</td>
<td>0.83</td>
<td>0.74</td>
<td>0.77</td>
<td>0.79</td>
<td>0.72</td>
<td>0.75</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ClinicalBERT</td>
<td>micro</td>
<td>0.88</td>
<td>0.88</td>
<td>0.88</td>
<td>0.75</td>
<td>0.75</td>
<td>0.75</td>
<td>0.83</td>
<td>0.83</td>
<td>0.83</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.93</td>
<td>0.93</td>
<td>0.93</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>macro</td>
<td>0.79</td>
<td>0.79</td>
<td>0.79</td>
<td>0.75</td>
<td>0.63</td>
<td>0.65</td>
<td>0.80</td>
<td>0.74</td>
<td>0.75</td>
<td>0.83</td>
<td>0.76</td>
<td>0.79</td>
<td>0.83</td>
<td>0.72</td>
<td>0.76</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Discussion

**Annotator disagreements**

The lower IAA in Certainty can be attributed to the unclear language in physician documentation, for example:

- “This was discussed by her neurologist who suggested starting with cholesterol medication and aspirin. She is here to discuss this further.”
- “We might suggest that she be started on Cisapride 10 mg qd”
- “It might be worthwhile to reduce his atenolol from 75 to 50mg once daily to see if this helps his Sx.”

Hedging language is common in clinical documentation to express the uncertainty inherent in medical decision-making. Reflecting this, different terms expressing varying degrees of certainty are used to indicate a prescribed intervention. For example, terms such as try, recommend, or advise are more definitive, while consider, suggest, or may benefit from require some interpretation to decide whether the action has actually taken place (Certain) or is only being discussed (Hypothetical). Although we attempted to address hedging language in the annotation guidelines, these disagreements reflect the inherent subjectivity in interpreting clinical language given the complexity and ambiguity in certain aspects of clinical documentation.

Disagreements along the Actor dimension were mostly found in references to past events, e.g. “Improved breathing with spiriva, mistakenly stopped Advair”, where the subject who initiated the action is unspecified and was interpreted...
differently by different annotators. Another interesting kind of disagreement was due to shared decision-making in clinical care, where both the patient and physician contribute to medical decisions and treatments, e.g. “Discussed changing from Avapro to losartan for cost issues but he is uninterested and does not want to rock the boat.”

**Error Analysis**

We conducted error analysis on the best performing model (i.e. ClinicalBERT) and identified three major categories of errors: (1) medication mentions with multiple annotations, (2) multiple medications within the same sentence, and (3) medication mentions that require context beyond the immediate sentence to determine the label. Examples for each of these error categories are shown in Table 4.

**Table 4:** Common error categories with examples, across five classification subtasks for the ClinicalBERT model.

<table>
<thead>
<tr>
<th>Error Category</th>
<th>Example</th>
<th>Medication</th>
<th>Ground Truth</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medication mentions with multiple annotations</td>
<td>In addition, 8 days prior to admission, pt’s regular lasix dose was increased from 80 to 120 mg for four days, then reduced back to 80 mg.</td>
<td>lasix</td>
<td>Increase</td>
<td>Decrease</td>
</tr>
<tr>
<td>Multiple medications within the same sentence</td>
<td>We could change his statin from Mevacor to Lipitor to increase the HDL.</td>
<td>Mevacor, Lipitor</td>
<td>Stop</td>
<td>Start</td>
</tr>
<tr>
<td>Limited context</td>
<td>P: He will restart glyburide 5 mg q.d. when his blood sugar is greater than 200.</td>
<td>glyburide</td>
<td>Conditional</td>
<td>Certain</td>
</tr>
</tbody>
</table>

A significant percentage of errors occur due to medications having multiple event annotations. For example, in “In addition, 8 days prior to admission, pt’s regular lasix dose was increased from 80 to 120 mg for four days, then reduced back to 80 mg.” the medication lasix has two labels for the Action dimension (Increase & Decrease). Since our current classification setup only allows for one prediction per medication mention, the model predicts only a single label Decrease for the Action dimension, leading to an error. Although medication mentions with multiple annotations form a small fraction of our overall dataset (1.2%), this error category accounts for a large percentage of errors across all dimensions (33% of Action errors, 23% for Temporality, 22% for Certainty, and 32% for Actor). One way to address this is to reformulate our task from a sentence classification task to a multi-label classification task.

Next, we observed that multiple medications present within the same sentence lead to errors as the model is unable to differentiate between the target medications. For example, “We could change his statin from Mevacor to Lipitor to increase the HDL.” contains two medications, Mevacor and Lipitor, that each have an Action label of Stop and Start, respectively. However, since they share the same context (i.e. sentence), the model predicts both the labels as Start. To resolve this, the system needs to more precisely identify the context for each medication mention. This can be achieved by reformulating this task as a named entity recognition task.

Finally, we observed a number of errors due to the limited context of a single sentence being available to the model for prediction. For example, in Table 4, the strike-through text “when his blood sugar is greater than 200” was not fed into ClinicalBERT. Hence the model made the correct prediction (Certainty: Certain) under the limited context given (“P: He will restart glyburide 5 mg q.d.”). While improved sentence segmentation will help this instance, a more general solution such as sequential sentence classification is likely to improve this error category.

**Limitations**

We acknowledge certain limitations to our work, specifically, those due to the nature of the underlying corpus and those that can be attributed to our annotation guidelines.

CMED is built on top of the corpus used in the 2014 i2b2/UTHealth Natural Language Processing shared task. Since this corpus was selected for the purposes of the 2014 i2b2 shared task and therefore focused heavily on diabetes and heart disease patients, it is not representative of a typical patient population. Further, the corpus is limited to a single data warehouse i.e. Partners HealthCare Electronic Medical Records. Reproduction of our work on more diverse corpora is needed to better understand the effectiveness and applicability of our schema.

The current task focuses primarily on the identification and classification of contextual information for medication
change events. Medication mentions that do not discuss change are all grouped under a single label, i.e. NoDisposition, including descriptions of medication status (e.g. “currently taking lisinopril”), explicit directions to continue an existing medication (e.g. “continue metformin”), documented allergies to medications (e.g. “sulfa (rash)”), and other incidental mentions of medications. Depending on the specific use case and application, there may be value in further teasing out these different types of NoDisposition events. Further, although the current schema captures coarse temporality information of medication change events, extraction of more specific temporal references (e.g. “at last visit”, “x 10 days”) is needed to place these events in a more precise point in time. Finally, there may be additional contextual information that could contribute to improved understanding of medication changes but was not included in our annotation schema, such as the magnitude of change (i.e. what is the degree of change?) and the reason behind the change (i.e. why was this change introduced?). Such information was excluded from our current effort because of difficulties in defining a set of discrete labels to capture all possible values. Future work can be undertaken to provide such information through an extraction task built on top of CMED.

Applications and Future Work

Medication change events classified under the proposed schema can be directly leveraged in several real-world applications. Various visualizations and dashboard displays have been proposed to improve the usability of EHR systems, many of which include a medication timeline based on structured medication data. Future research can be undertaken to apply analytics developed on CMED to such applications. For example, Present and Certain actions identified from clinical narratives under our schema can be incorporated into such medication timelines to further enrich them for a more comprehensive representation of a patient’s medication history. These same events can also be presented alongside structured medication data to surface potentially missed or incorrect medication information in the structured data for purposes of medication reconciliation. Separately, patient-initiated actions captured under our schema can be used to supplement pharmacy prescription filling data towards improved understanding of medication nonadherence. Figure 2 shows an example of how such extracted multi-dimensional medication events may be used at the point-of-care, allowing users to control the information flow depending on their needs and specific use case.

Figure 2: Prototype visualization incorporating extracted medication change events into structured EHR data showing all Present Certain events associated with the medication hydralazine.

CMED was purposefully generated on a longitudinal corpus (2014 i2b2 corpus with 2-5 notes per patient) to allow future work in tracking events for a specific medication over time, to support applications such as medication timeline visualization and summarization systems. To better support such applications and also more fully utilize the clinical context captured in CMED, additional research is necessary in several directions. First, 7.7% of medication mentions in CMED were classified as Undetermined, which can potentially be classified as either Disposition or NoDisposition.
tion given additional information. Further research in incorporating other sources of medication information (e.g. structured medication orders) is needed to better understand and characterize Undetermined events. Second, although temporal information is captured in the current schema, additional research is needed for applications that need temporal alignment or to place events in a more precise point in time. Third, the current schema classifies medications at an instance level, meaning that a single medication change occurrence that is documented multiple times in a patient’s medical record will result in multiple instance-level events. Therefore, further work in concept normalization and coreference resolution is needed to resolve instance-level events in and across clinical narratives. Finally, to make CMED more useful in real-world settings, more work needs to done to identify the contradictions between different events recorded for the same medication and resolve them.

Conclusion

We introduce CMED, a dataset capturing contextual information – Action, Negation, Temporality, Certainty, and Actor – for medication change events documented in clinical notes, consisting of 9,013 annotated medication mentions over 500 notes. We describe our annotation guidelines, discuss specific nuances observed during the annotation process, and explore state-of-the-art transformer-based models to automate the task. As the first dataset on medication change events to be made available to the research community, CMED provides the necessary first step towards improved understanding of medication events in clinical narratives. We hope this effort will encourage future research and exploration into leveraging medication information from clinical narratives, and also contribute to other use cases that require consideration of contextual information for clinical events.

References

Extraction of Electronic Health Record Data using Fast Healthcare Interoperability Resources for Automated Breast Cancer Risk Assessment

Julia E. McGuinness, MD1,2,3, Tianmai M. Zhang, MA1, Kevin Cooper BS4, Arusha Kelkar MS1, Jill Dimond PhD3, Virginia Lorenzi MS1, Katherine D. Crew MD MS2,3,5, Rita Kukafka, DrPH, MA, FACMI1

1Department of Biomedical Informatics, Vagelos College of Physicians and Surgeons, Columbia University, New York, NY, USA; 2 Department of Medicine, Vagelos College of Physicians and Surgeons, Columbia University, New York, NY, USA; 3Herbert Irving Comprehensive Cancer Center, Columbia University, New York, NY, USA; 4Sassafras Tech Collective, Ann Arbor, MI, USA; 5Department of Epidemiology, Mailman School of Public Health, Columbia University, New York, NY, USA

Abstract

Women at high risk for breast cancer may benefit from enhanced screening and risk-reduction strategies. However, limited time during clinical encounters is one barrier to routine breast cancer risk assessment. We evaluated if electronic health record (EHR) data downloaded using Fast Healthcare Interoperability Resources (FHIR) is sufficient for breast cancer risk calculation in our decision support tools, RealRisks and BNAV. We accessed EHR data using FHIR for six patient advocates, and downloaded and parsed XML documents. We searched for relevant clinical variables, and evaluated if data was sufficient to calculate risk using validated models (Gail, Breast Cancer Screening Consortium [BCSC], BRCAPRO). While only one advocate had sufficient EHR data to calculate risk using the BCSC model only, we identified variables including age, race/ethnicity, mammographic density, and prior breast biopsy in most advocates. EHR data from FHIR could be incorporated into automated breast cancer risk calculation in clinical decision support tools.

Introduction

Breast cancer is the most common cancer among women in the United States, with approximately 280,000 new cases and 40,000 deaths from breast cancer each year.1 While breast cancer mortality has declined over the past three decades because of improved detection and therapeutic advances, this decline has begun to plateau, particularly among racial/ethnic minorities.2 The identification of women who could benefit from enhanced breast cancer screening and preventive measures could further improve breast cancer mortality. Women at high risk for breast cancer, defined as an estimated 5-year risk of invasive breast cancer ≥1.67% and/or lifetime risk ≥20%, are eligible for chemoprevention with anti-estrogen agents, which reduce the risk of invasive breast cancer by approximately 50% to 65%.3,4 In addition, multiple national medical organizations recommend more intensive screening with annual mammography and consideration of supplemental breast MRI or ultrasound among high-risk women5-11. However, rates of chemoprevention uptake remain low12, with barriers that include concerns about potential side effects of treatment, insufficient knowledge about breast cancer risk assessment and chemoprevention among patients and providers, and time constraints during clinical encounters13-16. There is therefore a clear unmet need for interventions to facilitate accurate breast cancer risk assessment and enhance shared decision-making about breast cancer screening and risk-reducing measures.

We have developed the web-based decision support tools RealRisks and Breast cancer risk NAVIgation (BNAV) for patients and primary care providers, respectively, that calculate an individual patient’s estimated breast cancer risk using validated models17-18 and include interactive educational modules on breast cancer risk assessment, genetic testing, screening, and chemoprevention19-22. RealRisks is patient-facing and available in English and Spanish, while BNAV is provider-facing. Through these interventions, we aim to improve accuracy of patients’ breast cancer risk perception, enhance breast cancer screening and chemoprevention knowledge, and inform decisions regarding genetic testing, screening and risk reduction. One barrier to seamless integration of RealRisks and BNAV into clinic workflow is the requirement for manual input of patients’ personal and family history for individualized breast cancer risk calculations. Current breast cancer risk models incorporate information including patient age, race/ethnicity, breast density (assessed during mammography), reproductive history, prior breast biopsy (and pathologic diagnoses), and family history. Automation of patients’ breast cancer risk calculations using data extracted from the electronic health

843
record (EHR) could allow for more rapid, routine risk assessment. In addition, automated presentation of patients’ relevant medical histories to patients and providers could facilitate shared decision-making regarding therapeutic decisions such as recommendations regarding chemopreventive agents and lifestyle modification for risk reduction.

Fast Healthcare Interoperability Resources (FHIR) is a standard and application programming interface (API) created by the Health Level Seven International (HL7) health-care standards organization for the exchange of electronic healthcare information across platforms, particularly to support automated clinical decision support. To facilitate faster, logical, and consistent data exchange among healthcare applications, FHIR defines and organizes data elements into “resources,” such as “Condition,” “Procedure,” and “FamilyMemberHistory,” that are easily identified and understood by recipient applications. Multiple widely-used EHR vendors, including Epic and Cerner, support FHIR in order to comply with national regulatory requirements for interoperability from the Centers for Medicare and Medicaid Services (CMS) and Office of the National Coordinator for Health Information Technology (ONC). FHIR has previously been evaluated as a potential method of automatically populating documentation and registries in clinical trials, which might allow for streamlined identification of potential candidates for enrollment and more seamless clinical trial data sharing. FHIR therefore might also facilitate automated breast cancer risk calculation.

Utilizing a unique multidisciplinary team composed of biomedical informaticists, software developers, and medical oncologists from Columbia University Irving Medical Center in New York, NY, we evaluated whether the FHIR standard could support automated breast cancer risk calculations in RealRisks and BNAV as well as presentation of relevant patient medical history to patients and providers to facilitate shared decision-making.

Methods

Clinical Variables of Interest for Risk Calculation

Our primary objective was to find clinical variables within patient data downloaded from the FHIR API to allow for future automation of breast cancer risk calculation in RealRisks and BNAV. RealRisks and BNAV use two validated risk models, the Gail and Breast Cancer Screening Consortium models, to estimate a patient’s breast cancer risk, and the BRCAPRO tool to estimate a patient’s risk of carrying a deleterious mutation in BRCA1 and/or BRCA2, which are associated with up to an 80% lifetime risk of developing breast cancer. The clinical variables required for these models are summarized below (Table 1). While there are differences across models in the types of variables required, variables generally fall into one of five categories: 1) Demographics, 2) Reproductive factors, 3) Family history of breast cancer, 4) Breast pathology, and 5) Breast imaging.

Table 1. Summary of variables required for breast cancer risk calculation according to the Gail, Breast Cancer Surveillance Consortium (BCSC), and BRCAPRO models.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Gail</th>
<th>BCSC</th>
<th>BRCAPRO</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Demographics</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Race/Ethnicity</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td><strong>Reproductive Factors</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age at first live birth</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age at first menstrual period</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Family Cancer History</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First-degree relative with breast cancer</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Second-degree relative with breast cancer</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Relative with ovarian cancer</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Relative with male breast cancer</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Relative with bilateral breast cancer</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ashkenazi Jewish descent</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Breast Pathology</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>History of benign breast biopsy</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td># of breast biopsies</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>History of atypical hyperplasia</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Detailed breast pathology results</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td><strong>Breast Imaging</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mammographic density (BI-RADS)</td>
<td></td>
<td></td>
<td>X</td>
</tr>
</tbody>
</table>
Of note, the BCSC model requires classification of pathology results into: 1) Prior biopsy, unknown diagnosis; 2) Non-proliferative lesion; 3) Proliferative changes without atypia; 5) Proliferative changes with atypia; 6) Lobular carcinoma in situ, as well as classification of mammographic density into the four BI-RADS categories: 1) Almost entirely fatty; 2) Scattered fibroglandular densities; 3) Heterogeneously dense; 4) Extremely dense.

In addition, a secondary objective was to find additional relevant medical history within patient data downloaded from the FHIR API that could be presented within RealRisks and BNAV with the goal of facilitating shared decision-making about chemoprevention and other risk-aligned preventive options. The variables of interest are summarized below (Table 2). These variables were chosen for their potential to guide choice of chemoprevention (for example, menopausal status, history of thrombosis or endometrial cancer, osteoporosis), discussion of testing for hereditary cancer syndromes (family history of non-breast cancers including pancreatic and prostate), and recommendation for other non-chemoprevention risk-reducing measures (for example, alcohol use, body mass index [BMI]).

### Table 2. Summary of variables of interest to facilitate shared decision-making about breast cancer prevention.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Personal Medical and Surgical History</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>History of venous or arterial thrombosis, pulmonary embolism</td>
</tr>
<tr>
<td></td>
<td>Cardiovascular risk factors: Hypertension, Hyperlipidemia, Diabetes mellitus, Coronary artery disease, Congestive heart failure</td>
</tr>
<tr>
<td></td>
<td>Osteoporosis or osteopenia</td>
</tr>
<tr>
<td></td>
<td>Endometrial cancer</td>
</tr>
<tr>
<td></td>
<td>Hysterectomy and/or oophorectomy</td>
</tr>
<tr>
<td></td>
<td>Bilateral mastectomy</td>
</tr>
<tr>
<td></td>
<td>Gynecologic History</td>
</tr>
<tr>
<td></td>
<td>Menopausal status (premenopausal, perimenopausal, postmenopausal)</td>
</tr>
<tr>
<td></td>
<td>Age at menopause</td>
</tr>
<tr>
<td></td>
<td>Number of live births</td>
</tr>
<tr>
<td></td>
<td>Use of oral contraceptives</td>
</tr>
<tr>
<td></td>
<td>Use of hormone replacement therapy</td>
</tr>
<tr>
<td></td>
<td>Social History</td>
</tr>
<tr>
<td></td>
<td>Alcohol use</td>
</tr>
<tr>
<td></td>
<td>Smoking history</td>
</tr>
<tr>
<td></td>
<td>Drug use</td>
</tr>
<tr>
<td></td>
<td>Vital Signs</td>
</tr>
<tr>
<td></td>
<td>Body mass index (kg/m²)</td>
</tr>
<tr>
<td></td>
<td>Systolic and diastolic blood pressure</td>
</tr>
<tr>
<td></td>
<td>Medications (Current or Previous)</td>
</tr>
<tr>
<td></td>
<td>Chemoprevention (tamoxifen, raloxifene, anastrozole, exemestane)</td>
</tr>
<tr>
<td></td>
<td>Oral contraceptive pills</td>
</tr>
<tr>
<td></td>
<td>Estrogen- and/or progesterone-containing medications</td>
</tr>
<tr>
<td></td>
<td>Family Cancer History</td>
</tr>
<tr>
<td></td>
<td>Relative with pancreatic cancer</td>
</tr>
<tr>
<td></td>
<td>Relative with prostate cancer</td>
</tr>
<tr>
<td></td>
<td>Relative with colorectal cancer</td>
</tr>
<tr>
<td></td>
<td>Relative with endometrial cancer</td>
</tr>
<tr>
<td></td>
<td>Imaging</td>
</tr>
<tr>
<td></td>
<td>Breast magnetic resonance imaging (MRI)</td>
</tr>
<tr>
<td></td>
<td>Breast ultrasound</td>
</tr>
<tr>
<td></td>
<td>Bone density scan (DEXA)</td>
</tr>
<tr>
<td></td>
<td>Genetic Testing/Counseling</td>
</tr>
<tr>
<td></td>
<td>Genetic counseling notes</td>
</tr>
<tr>
<td></td>
<td>Germline genetic testing results (BRCA1/2, multigene panel testing)</td>
</tr>
<tr>
<td></td>
<td>Lab Tests</td>
</tr>
<tr>
<td></td>
<td>FSH, LH, estradiol (to assess menopausal status)</td>
</tr>
</tbody>
</table>
We invited six patient advocates to access RealRisks and allow for their EHR data to be downloaded using our RealRisks SMART on FHIR add-on. We used FHIR versions Second Draft Standard for Trial Use (DSTU2), Release 3 Standard for Trial Use (STU3), and Release 4 (R4), and only accessed data from medical providers utilizing the Epic EHR. In order to locate clinical variables within FHIR, we downloaded data from the following FHIR resources: “Patient,” “CarePlan,” “DiagnosticReport,” “Condition,” “MedicationStatement,” “Procedure,” “Observation” (Smoking History, Vitals, Labs), “FamilyMemberHistory,” and “DocumentReference” (Consolidated-Clinical Document Architecture [CCDA] documents). These resources consisted of documents formatted via XML.

### Data Parsing

We automatically parsed patient birthdate, race, and ethnicity from the FHIR “Patient” resource. The remainder of data from the FHIR API was downloaded the form of XML documents, which we parsed to have a tree structure using the lxml package (version 4.5.2) in Python (version 3.8.3). Two file types, “Patient Health Summary” and “Encounter Summary,” as well as many components contained in these files (such as patient allergies, medications, problems, test results, and coded diagnoses) were identified and separated for information extraction. Keywords related to the variables of interest were searched using regular expression and then exported.

### Results

We successfully accessed EHR data for all six patient advocates. We automatically populated patient birthdate, race and ethnicity from the FHIR “Patient” resource into the RealRisks application. The RealRisks patient-facing window for risk calculation using the BCSC model (Figure 1) incorporates automatically populated variables alongside variables that were manually populated by the study team at time of patient RealRisks account creation.

![Figure 1](image-url)  
*Figure 1.* Example of a RealRisks patient-facing window for risk calculation using the BCSC model. Birthdate, race, and ethnicity (on the left) are automatically populated from the FHIR “Patient” resource. Family history of breast cancer in a first-degree relative was not auto-populated, and was documented as “unknown.” When family history is not available in the EHR, patients can enter this data by interacting with the pedigree generating function in RealRisks. On the right, variables including personal history of breast cancer or surgery, mammographic density, and prior breast biopsy were manually populated by the study team at the time of RealRisks account creation, and the patient is given the option to request changes in the current version.

We then parsed the “Patient Health Summary” and “Encounter Summary” XML files for each of the six advocates. While each patient had one “Patient Health Summary” file summarizing key clinical data, a single patient could have multiple “Encounter Summary” documents, with each file representing a single encounter (such as office visit, hospitalization, imaging study). For example, one patient advocate had 153 “Encounter Summary” files representing 153 separate encounters in Epic. The contents of these files are summarized below (Table 3).
Table 3. Contents of “Patient Health Summary” and “Encounter Summary” XML files downloaded from FHIR. A component would be present in the document if the patient has relevant records.

<table>
<thead>
<tr>
<th>Component Name</th>
<th>LOINC Code</th>
<th>Description of Contents</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Patient Health Summary</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patient information</td>
<td></td>
<td>Patient name, gender, birth date, marital status, religion, race, ethnicity, language.</td>
</tr>
<tr>
<td>Allergies</td>
<td>48765-2:</td>
<td>Allergies and adverse reactions Document</td>
</tr>
<tr>
<td></td>
<td>Allergies and adverse reactions Document</td>
<td></td>
</tr>
<tr>
<td>Medications</td>
<td>10160-0:</td>
<td>History of Medication use Narrative</td>
</tr>
<tr>
<td></td>
<td>Active medications, medication start time, doses, administration time, and administration methods. Drug codes in but not limited to RxNorm.</td>
<td></td>
</tr>
<tr>
<td>Active Problems</td>
<td>11450-4:</td>
<td>Problem - Reported</td>
</tr>
<tr>
<td></td>
<td>Active problems and time. Codes in SNOMED-CT, ICD9, ICD10, and Intelligent Medical Objects ProblemIT.</td>
<td></td>
</tr>
<tr>
<td>Resolved Problems</td>
<td>11348-0:</td>
<td>History of past illness</td>
</tr>
<tr>
<td></td>
<td>Resolved problems and time. Codes in SNOMED-CT, ICD9, ICD10, and Intelligent Medical Objects ProblemIT.</td>
<td></td>
</tr>
<tr>
<td>Immunizations</td>
<td>11369-6:</td>
<td>History of Immunization Narrative</td>
</tr>
<tr>
<td></td>
<td>Immunization names, time, manufacturers, and performers. Substance codes in CVX.</td>
<td></td>
</tr>
<tr>
<td>Procedures</td>
<td>47519-4:</td>
<td>History of Procedures Document</td>
</tr>
<tr>
<td></td>
<td>Procedure names, time, reasons of performing, and performer. Procedure codes in Epic.EAP.ID.</td>
<td></td>
</tr>
<tr>
<td>Results</td>
<td>30954-2:</td>
<td>Relevant diagnostic tests/laboratory data Narrative</td>
</tr>
<tr>
<td></td>
<td>Results of laboratory tests. Sorted by time. Up to 200 results per patient.</td>
<td></td>
</tr>
<tr>
<td><strong>Encounter Summary</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patient information</td>
<td></td>
<td>Same as above</td>
</tr>
<tr>
<td>Reason for Visit</td>
<td>29299-5:</td>
<td>Reason for visit, specialty, diagnosis/procedures, provider referred (by and to).</td>
</tr>
<tr>
<td></td>
<td>8661-1:</td>
<td>Chief Complaint</td>
</tr>
<tr>
<td>Encounter Details</td>
<td>46240-8:</td>
<td>History of encounters</td>
</tr>
<tr>
<td></td>
<td>Encounter date, type, department, location, care team.</td>
<td></td>
</tr>
<tr>
<td>Social History</td>
<td>29762-2:</td>
<td>Social history Narrative</td>
</tr>
<tr>
<td></td>
<td>Tobacco use (including packs/day, years used), alcohol use (including drinks/week, oz/week), sex assigned at birth, occupation (including job start date, industry), travel history. Codes in SNOMED-CT and LOINC.</td>
<td></td>
</tr>
<tr>
<td>Last Filed Vital Signs</td>
<td>8716-3:</td>
<td>Vital signs</td>
</tr>
<tr>
<td></td>
<td>Blood pressure, pulse, temperature, respiratory rate, oxygen saturation, height, weight, body mass index.</td>
<td></td>
</tr>
<tr>
<td>Progress Notes</td>
<td>10164-2:</td>
<td>History of Present Illness</td>
</tr>
<tr>
<td></td>
<td>Sharing of progress notes is not supported in DSTU2.</td>
<td></td>
</tr>
<tr>
<td>Anesthesia Record</td>
<td>59774-0:</td>
<td>Anesthesia Records</td>
</tr>
<tr>
<td></td>
<td>Anesthesia records, date, provider.</td>
<td></td>
</tr>
<tr>
<td>Medications at Time of Discharge</td>
<td>10183-2:</td>
<td>Hospital Discharge Medications</td>
</tr>
<tr>
<td></td>
<td>Medications, Signatura, amount dispensed, date, and author. Drug codes in but not limited to RxNorm.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>75311-1:</td>
<td>Discharge Medications</td>
</tr>
<tr>
<td>Plan of Treatment</td>
<td>18776-5:</td>
<td>Plan of care note</td>
</tr>
<tr>
<td></td>
<td>Date, type, specialty, and care team of upcoming encounter.</td>
<td></td>
</tr>
<tr>
<td>Patient Instructions</td>
<td>61146-7:</td>
<td>Instructions</td>
</tr>
<tr>
<td></td>
<td>Provider instructions during the encounter.</td>
<td></td>
</tr>
<tr>
<td>Discharge Instructions</td>
<td>8653-8:</td>
<td>Discharge Instructions</td>
</tr>
<tr>
<td></td>
<td>Provider instructions at discharge and patient education.</td>
<td></td>
</tr>
<tr>
<td>Goals</td>
<td>61146-7:</td>
<td>Goals</td>
</tr>
<tr>
<td></td>
<td>Goals of the encounter.</td>
<td></td>
</tr>
<tr>
<td>Procedures</td>
<td>47519-4:</td>
<td>History of Procedures Document</td>
</tr>
<tr>
<td></td>
<td>Relevant procedures.</td>
<td></td>
</tr>
<tr>
<td>Results</td>
<td>30954-2:</td>
<td>Relevant diagnostic tests/laboratory data Narrative</td>
</tr>
<tr>
<td></td>
<td>Relevant test results.</td>
<td></td>
</tr>
<tr>
<td>Visit Diagnoses</td>
<td>51848-0:</td>
<td>Assessments</td>
</tr>
<tr>
<td></td>
<td>Diagnoses associated with the encounter.</td>
<td></td>
</tr>
<tr>
<td>Administered Medications</td>
<td>29549-3:</td>
<td>Medications administered</td>
</tr>
<tr>
<td></td>
<td>Medications, Signatures, action dates, and author.</td>
<td></td>
</tr>
</tbody>
</table>
Shown below are available FHIR data for breast cancer risk calculation and additional relevant clinical variables for each patient advocate and in total (Table 4). Among those variables required for risk calculation, variables with documentation in the majority of advocates were age (100%), race/ethnicity (67%), history of benign breast biopsy and number of breast biopsies (67%), and mammographic breast density (67%). Only one-third of advocates had documentation of family history including first- or second-degree relative with breast cancer and relative with ovarian cancer, male breast cancer, or bilateral breast cancer, and none had documentation of Ashkenazi descent or reproductive factors including age at menarche and age at first live birth. While one (17%) had documentation of a history of atypical hyperplasia, no patient advocate had available breast pathology reports. One advocate (Patient Advocate #1) had sufficient data from the FHIR XML files to calculate risk using the BCSC model, but no patient advocate had sufficient data to calculate risk using the Gail model or BRCAPRO.

Among relevant (but not required) clinical variables, all advocates (100%) had documentation of blood pressure as well as height and weight for calculation of BMI. Half of advocates had breast ultrasound reports, and one had a breast MRI report. The remainder of variables had documentation in one or no advocate. For example, while one advocate had documentation of menopausal status, none had age of menopause or number of live births documented. No advocate had a genetic testing results report or a genetic counseling note.

Table 4. Available FHIR data for relevant clinical variables for each patient advocate and in total. An X denotes that documentation of a variable was found in XML files for that individual “Continued”.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Patient Advocate</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Necessary for Risk Calculation</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>X X X X X X X X</td>
<td>100% (6/6)</td>
</tr>
<tr>
<td>Race/Ethnicity</td>
<td>X X X X</td>
<td>67% (4/6)</td>
</tr>
<tr>
<td>History of benign breast biopsy</td>
<td>X X X</td>
<td>67% (4/6)</td>
</tr>
<tr>
<td>Number of breast biopsies</td>
<td>X X X</td>
<td>67% (4/6)</td>
</tr>
<tr>
<td>Mammographic density</td>
<td>X X X</td>
<td>67% (4/6)</td>
</tr>
<tr>
<td>First-degree relative with breast cancer</td>
<td>X X X</td>
<td>33% (2/6)</td>
</tr>
<tr>
<td>Second-degree relative with breast cancer</td>
<td>X X X</td>
<td>33% (2/6)</td>
</tr>
<tr>
<td>Relative with ovarian cancer, male breast cancer, or bilateral breast cancer</td>
<td>X X X</td>
<td>33% (2/6)</td>
</tr>
<tr>
<td>History of atypical hyperplasia</td>
<td>X X X</td>
<td>33% (2/6)</td>
</tr>
<tr>
<td>Breast pathology reports available</td>
<td>X X X</td>
<td>33% (2/6)</td>
</tr>
<tr>
<td>Age at first menstrual period</td>
<td>X X X</td>
<td>33% (2/6)</td>
</tr>
<tr>
<td>Age at first live birth</td>
<td>X X X</td>
<td>33% (2/6)</td>
</tr>
<tr>
<td>Ashkenazi Jewish descent</td>
<td>X X X</td>
<td>33% (2/6)</td>
</tr>
<tr>
<td><strong>Additional Relevant Variables</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Body mass index (kg/m²)</td>
<td>X X X X X X X</td>
<td>100% (6/6)</td>
</tr>
<tr>
<td>Systolic and diastolic blood pressure</td>
<td>X X X X X X X</td>
<td>100% (6/6)</td>
</tr>
<tr>
<td>Alcohol use</td>
<td>X X X</td>
<td>50% (3/6)</td>
</tr>
<tr>
<td>Breast ultrasound</td>
<td>X X X</td>
<td>50% (3/6)</td>
</tr>
<tr>
<td>Chemoprevention use</td>
<td>X X X</td>
<td>33% (2/6)</td>
</tr>
<tr>
<td>Use of oral contraceptives</td>
<td>X X X</td>
<td>33% (2/6)</td>
</tr>
<tr>
<td>Smoking history</td>
<td>X X X</td>
<td>33% (2/6)</td>
</tr>
<tr>
<td>Cardiovascular risk factors</td>
<td>X X X</td>
<td>33% (2/6)</td>
</tr>
<tr>
<td>Bilateral mastectomy</td>
<td>X X X</td>
<td>33% (2/6)</td>
</tr>
<tr>
<td>Menopausal status</td>
<td>X X X</td>
<td>33% (2/6)</td>
</tr>
<tr>
<td>Family history of non-breast cancers</td>
<td>X X X</td>
<td>33% (2/6)</td>
</tr>
<tr>
<td>Breast magnetic resonance imaging (MRI)</td>
<td>X X X</td>
<td>33% (2/6)</td>
</tr>
<tr>
<td>Bone density scan (DEXA)</td>
<td>X X X</td>
<td>33% (2/6)</td>
</tr>
<tr>
<td>FSH, LH, estradiol</td>
<td>X X X</td>
<td>33% (2/6)</td>
</tr>
<tr>
<td>Age at menopause</td>
<td>X X X</td>
<td>33% (2/6)</td>
</tr>
<tr>
<td>Number of live births</td>
<td>X X X</td>
<td>33% (2/6)</td>
</tr>
<tr>
<td>Use of hormone replacement therapy</td>
<td>X X X</td>
<td>33% (2/6)</td>
</tr>
</tbody>
</table>
Table:

<table>
<thead>
<tr>
<th>Condition</th>
<th>RealRisks</th>
<th>BNAV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hysterectomy and/or oophorectomy</td>
<td>0% (0/6)</td>
<td></td>
</tr>
<tr>
<td>Osteoporosis or osteopenia</td>
<td>0% (0/6)</td>
<td></td>
</tr>
<tr>
<td>Endometrial cancer</td>
<td>0% (0/6)</td>
<td></td>
</tr>
<tr>
<td>History of venous or arterial thrombosis, pulmonary embolism</td>
<td>0% (0/6)</td>
<td></td>
</tr>
<tr>
<td>Drug use</td>
<td>0% (0/6)</td>
<td></td>
</tr>
<tr>
<td>Genetic counseling notes</td>
<td>0% (0/6)</td>
<td></td>
</tr>
<tr>
<td>Germline genetic testing reports</td>
<td>0% (0/6)</td>
<td></td>
</tr>
</tbody>
</table>

Discussion:

We evaluated whether the FHIR standard could support automated breast cancer risk calculations in RealRisks using data downloaded for six patient advocates. While only one patient advocate had sufficient information for risk calculation using the BCSC model and none had complete information for risk calculation using the Gail or BRCAPRO models, we found that the majority of advocates had documentation of age, race/ethnicity, prior breast biopsy, and mammographic density. We also found additional data, including vital signs (blood pressure, BMI) and coexisting medical conditions, that would be relevant to patient-provider discussions and decision-making regarding chemoprevention and risk-reducing measures.

To our knowledge, our group is the first to investigate the use of the FHIR API to enhance and automate breast cancer risk calculation. Our current project represents our initial attempts to locate relevant patient data in downloads from the FHIR API, and even with a small sample size of only six patients, demonstrates that data for variables including breast pathology, breast imaging, vital signs, and race/ethnicity can be successfully found in FHIR downloads. We previously demonstrated that there was moderate agreement between information on breast cancer risk extracted from the EHR compared to patient self-report30, and also found that while self-report identifies more women who are potentially eligible for chemoprevention or genetic testing for hereditary breast and ovarian cancer (HBOC) based upon family history, more specific information on breast pathology or mammographic density can be obtained from the EHR30,31.

In this study, certain categories of patient data, particularly gynecologic history, family history of cancer, and history of genetic counseling and testing, were documented less frequently than other data types or not at all. This is consistent with our previous finding that family history is better represented in self-report than EHR data, but also highlights the limitations of using EHR data alone in risk calculation, with the potential for missing or inaccurate data12,34. The presence of information is dependent on documentation by providers in the EHR, which often does not include regular documentation of pertinent negative history (for example, that a patient does not have a family history of breast cancer) and also varies based upon individual providers’ scope of practice. Some missing data can also be explained by current gaps in FHIR coverage. For example, FHIR DSTU2 does not support sharing progress notes, and gynecologic history and family history are often present free-text in progress notes and not in structured data. FHIR R4 is anticipated to provide this coverage, but R4 support for progress notes was not available at the time of our study. However, an important consideration is that much of the information required for adequate and accurate risk assessment, if present, is dispersed throughout patients’ medical charts, recorded by multiple providers in different places such as progress notes and mammography reports and in both structured and unstructured form35. Finding such data requires approaches that search for free-text within notes and other documents, in addition to structured data such as billing codes.

One potential solution for automated risk calculation in our RealRisks and BNAV decision support tools is to incorporate both self-reported data and data automatically populated using the FHIR API, with the goal of harnessing the strengths of each approach. We plan to update RealRisks to automatically populate the necessary patient information for risk calculation, with a screen that shows this data to patients with a request to review and modify data before running the risk assessments. This will allow for patients to add any missing data and correct erroneous data, with the goal of making the process faster and more accurate. If feasible, an ultimate goal is to store this semi-automated, patient-corrected risk information and return to the EHR for ongoing use by providers; given that our work is still in early stages, we have not yet implemented this. In future, we also plan to develop a screen for both patients and providers to review automatically populated additional medical history that would be relevant to shared decision-making about breast cancer risk reduction with chemoprevention, testing for susceptibility genes, screening and other preventive strategies. While a personalized approach to prevention or early detection of breast cancer has emerged as a highly promising strategy, several barriers exist including the time required to conduct risk assessment of each women in a population36. Presenting a summary of auto-populated relevant history to the provider could save
time during patient encounters and enhance discussions and recommendations about personalized screening and prevention options, which could overcome some barriers to increasing the uptake of underutilized breast cancer prevention services.

We are currently evaluating RealRisks and BNAV in a randomized controlled trial among women with high-risk breast lesions (atypical hyperplasia, lobular or ductal carcinoma in situ); patients and their providers will be randomized to have access to these tools plus standard educational materials versus standard educational materials alone, and the frequency of chemoprevention informed choice will be compared at six months[77]. Given that the majority of participating sites in this trial use Epic, we anticipate that we will have access through the RealRisks SMART on FHIR add-on to many participants’ EHR data, and therefore we plan to implement our automated risk calculation method described above to evaluate usability among a larger cohort of women. In particular, this trial will allow us to evaluate the accuracy of the EHR data pulled from FHIR through comparison with breast cancer risk data that is manually entered by the study team at time of patient enrollment. In addition, our data parsing during this initial study was more time-intensive given that we sought to locate data within the downloaded data files, but we will next use natural language-processing methods to parse unstructured data from progress notes, particularly to automatically extract family history.

Strengths of our study, as stated above, include its innovative approach to meet the unmet clinical need of automating breast cancer risk calculations utilizing FHIR data, as well as our creation of a multidisciplinary team including medical oncologists, biomedical informaticists, and software developers to best inform our study questions and approach. A major limitation is its small sample size of only six patients, of whom all were patient advocates known to the study team; this likely introduced bias into the type of data variables that would be found (or missing) in the downloaded files, and therefore might limit the generalizability of our findings. However, this study represented an initial analysis of the feasibility of using FHIR to obtain breast cancer risk factors from the EHR, and further evaluation of our methods using a larger population in a prospective clinical trial will overcome these limitations.

**Conclusion:**

We were able to identify relevant patient data in FHIR that could be incorporated into automated breast cancer risk calculation in the RealRisks and BNAV decision support tools. We will next evaluate an automated risk calculation method incorporating patient EHR data from FHIR as part of an ongoing prospective clinical trial among patients at high risk for developing breast cancer.

**Acknowledgements:**

This work was supported by the National Institutes of Health (NIH), National Cancer Institute (NCI) R01CA177995 R01CA226060 (S1904), P30 CA013696; an American Cancer Society (ACS) Research Scholar Grant RS G-17-103-01; and a National Library of Medicine Biomedical Informatics Training Award T15 LM007079-29. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

**References**


On Predicting Recurrence in Early Stage Non-small Cell Lung Cancer

Sameh K. Mohamed1,2, Brian Walsh1,2, Mohan Timilsina1,2, Maria Torrente6, Fabio Franco6, Mariano Provencio6, Adrianna Janik3, Luca Costabello4, Pasquale Minervini4, Pontus Stenetorp4, Vít Nováček1,5

1Data Science Institute, NUI Galway, Galway, Ireland
2Insight Centre for Data Analytics, NUI Galway, Galway, Ireland
3Accenture Labs, Dublin, Ireland
4University College London, London, United Kingdom
5Faculty of Informatics, Masaryk University, Brno, Czech Republic
6Medical Oncology Department, Hospital Universitario Puerta de Hierro Majadahonda, Madrid, Spain

Abstract

Early detection and mitigation of disease recurrence in non-small cell lung cancer (NSCLC) patients is a nontrivial problem that is typically addressed either by rather generic follow-up screening guidelines, self-reporting, simple nomograms, or by models that predict relapse risk in individual patients using statistical analysis of retrospective data. We posit that machine learning models trained on patient data can provide an alternative approach that allows for more efficient development of many complementary models at once, superior accuracy, less dependency on the data collection protocols and increased support for explainability of the predictions. In this preliminary study, we describe an experimental suite of various machine learning models applied on a patient cohort of 2442 early stage NSCLC patients. We discuss the promising results achieved, as well as the lessons we learned while developing this baseline for further, more advanced studies in this area.

Introduction

Lung cancer is the most common cause of cancer related mortality in the world. Globally in 2018, more than 2 million new cases were diagnosed and more than 1.8 million deaths due to lung cancer occurred1, where the main risk factor for the development of the disease is tobacco. In recent years, while mortality in men has decreased slightly due to smoking cessation, mortality in women has increased and has practically doubled due to its later incorporation into smoking habit2. Non-small cell lung cancer (NSCLC) accounts for approximately 80% of all lung malignancies. The diagnosis of lung cancer in early stages remains a challenge and often occurs incidentally in the study of other diseases3. Approximately 70% of patients are diagnosed in locally advanced stages (stage III) or metastatic disease (stage IV)4, which contributes to low survival rates. Of note, survival declines progressively with increasing clinical stage. Overall survival at 5 years in NSCLC is around 10–15%5. Early stage NSCLC (stage I-II) patients are typically treated with complete surgical resection of the tumor. However, even after the entire resection of the tumor, 30–55% of patients will develop disease recurrence within the first 5 years of surgery6. This encouraged research into applying survival analysis methods and machine learning models to identify patients with high risk for recurrence to help in personalizing surveillance plans for these patients.

Predictive approaches for identifying tumor recurrence in NSCLC patients utilize different types of techniques such as statistical analysis of patient Electronic Health Records (EHRs), machine learning models and patient self-assessment feedbacks. The simplest form of these methods is the self-assessment techniques which rely on continuous patient self-evaluation and reporting after successful treatment7. This technique traditionally requires patients to perform self-assessment in terms of symptoms, physical abilities and other metrics, and report these metrics to hospitals through previously specified means e.g. mobile applications, and online forms8. It then notifies the physician when certain patterns appear in the patient recorded metrics8. Due to the simplicity of the approach, it cannot provide early predictions of the tumor recurrence, and it relies on patients compliance to reporting which is not guaranteed9.

On the other hand, statistical survival analysis methods and supervised machine learning models are able to provide early predictions for the tumor recurrence based on the patient characteristics and previous treatments history. One of the simplest forms of the statistical analysis tools is nomograms10, which are a graphical representations of equations that predict medical outcome. Nomograms use a points-based system whereby a patient accumulates points based on levels of his or her risk factors, where the cumulative points total is associated with a prediction. More complex techniques such as the Cox hazard model can provide a time-based analysis of the patient hazards e.g. tumor recurrence.
The Cox hazard model is one of the most widely used techniques in predicting recurrence and other prognosis factors in NSCLC patients. Despite its popularity, however, the Cox hazard model operates strictly with multiple assumptions, e.g., feature independence, linear relations, etc., which are not always guaranteed. It also cannot consume heterogeneous data such as medical imagery, genetic arrays, etc. These two limitations are addressed in supervised machine learning models such as random forests, support vector machines, neural networks, etc., which operate using different techniques that make them suitable for modeling complex relations on heterogeneous data.

Supervised models were used in different studies to predict NSCLC recurrence from patients' EHRs, where different studies have shown that they provide superior predictive accuracy compared to classical survival analysis methods, such as the Cox hazard model. However, these studies traditionally only examine a limited set of supervised machine learning approaches, and they also rely on differently structured patient records due to the different data collection protocols. These two factors make it difficult to perform a more representative comparison between the survival analysis methods and the different supervised machine learning models. Therefore, it is essential to establish rigorous means for comparison between the different supervised predictive models and statistical survival analysis techniques on unified patient EHRs structure to understand the differences between these techniques, and establish informative baselines for the task of predicting NSCLC recurrence.

In this work, we study the problem of predicting recurrence in NSCLC patients where we discuss our first-steps towards building an efficient and accurate prediction approach within the CLARIFY project — a European project focused on monitoring health status and quality of life after the cancer treatment (cf., https://www.clarify2020.eu/). Based on the previously discussed developments and challenges in relation to predicting recurrence in NSCLC patients, we started our development by assessing the performance of the basic supervised machine learning models in NSCLC recurrence prediction to establish a baseline benchmark.

In this study, we provide the outcomes of our assessment of the basic supervised machine learning methods on patient cohort of 2442 early stage NSCLC patients, where we evaluate these models on a binary classification task with the objective of classifying patients with successful treatments into one of two categories: tumor recurrence or disease-free survival. We also discuss the challenges associated to the application of these methods in terms of model accuracy, data quality, and utilized evaluation protocols.

Related Work

In the following, we discuss other works related to our study where we categorize these works into two categories: survival analysis methods and other machine learning supervised and unsupervised methods.

- **Survival analysis methods.** Survival analysis methods such as the Kaplan–Meier and Nelson–Aalen estimators, the proportional hazards models, etc., are the most common approach for predicting prognosis, relapse and other clinical outcomes of lung cancer patients. For example, (author?) used the Cox hazard to predict long-term mortality/survival after lung cancer resection in patients older than 65 years on Medicare data for lung cancer resections. These models were also used in more specific survival analysis studies which focused on patients' survival in relation to specific bio-markers and chemotherapy drug configurations. Similarly, study conducted by Wen et.al. in 393 North American patients with NSCLC used Cox proportional hazard model to investigate associations between functional genetic variants of autophagy-related genes and radiation pneumonitis as well as clinical outcomes after definitive radiotherapy. The multi-variable Cox model enabled them to predict radiation pneumonitis, local recurrence-free survival, progression-free survival, and overall survival of NSLC patients.

On a similar note, (author?) used survival analysis methods in their study to examine the causes of death of long-term survivors of lung cancer where they utilized the Surveillance, Epidemiology and End Results (SEER) database (1988–2008). They conducted a survival analysis using a Kaplan–Meier estimator and conducted a multivariate analysis using a Cox proportional hazard where the study concluded that cardiac as well as non-malignant pulmonary condition contributes considerable proportion of deaths in long-term lung cancer survivors.

- **Supervised machine learning models.** Multiple studies of non-small cell lung cancer (NSCLC) postoperative recurrence have utilized supervised machine learning model to predict the probability of tumour recurrence and other
### Table 1: Analysis of the characteristics of the patient cohort examined in our study in relation to tumor recurrence. The summation of the total percentages of the sub characteristics can be less than 100 for some of the examined features due to missing data values of this type of data for some patients EHRs.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Recurrence</th>
<th>Survival</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1275 (52.2%)</td>
<td>1167 (47.8%)</td>
<td>2442 (100.0%)</td>
</tr>
<tr>
<td>Age</td>
<td>65.9 (25-89)</td>
<td>65.4 (26-117)</td>
<td>65.6 (25-117)</td>
</tr>
<tr>
<td>Gender</td>
<td>Male</td>
<td>Female</td>
<td></td>
</tr>
<tr>
<td></td>
<td>989 (77.6%)</td>
<td>286 (22.4%)</td>
<td>1275 (52.2%)</td>
</tr>
<tr>
<td></td>
<td>861 (73.8%)</td>
<td>306 (26.2%)</td>
<td>1167 (47.8%)</td>
</tr>
<tr>
<td></td>
<td>1850 (75.8%)</td>
<td>592 (24.2%)</td>
<td>2442 (100.0%)</td>
</tr>
<tr>
<td>Smoking history</td>
<td>Current/Previous</td>
<td>Non smoker</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1115 (87.5%)</td>
<td>160 (12.5%)</td>
<td>2275 (93.0%)</td>
</tr>
<tr>
<td></td>
<td>1000 (85.7%)</td>
<td>167 (14.3%)</td>
<td>2167 (88.7%)</td>
</tr>
<tr>
<td></td>
<td>2115 (86.6%)</td>
<td>327 (13.4%)</td>
<td>2442 (100.0%)</td>
</tr>
<tr>
<td>Cancer stage</td>
<td>I</td>
<td>IA</td>
<td>IB</td>
</tr>
<tr>
<td></td>
<td>9 (0.7%)</td>
<td>31 (2.7%)</td>
<td>40 (1.6%)</td>
</tr>
<tr>
<td></td>
<td>289 (22.7%)</td>
<td>353 (30.2%)</td>
<td>642 (26.3%)</td>
</tr>
<tr>
<td></td>
<td>348 (27.3%)</td>
<td>303 (26.0%)</td>
<td>651 (26.7%)</td>
</tr>
<tr>
<td></td>
<td>18 (1.4%)</td>
<td>8 (0.7%)</td>
<td>26 (1.1%)</td>
</tr>
<tr>
<td></td>
<td>246 (19.3%)</td>
<td>177 (15.2%)</td>
<td>423 (17.3%)</td>
</tr>
<tr>
<td></td>
<td>365 (28.6%)</td>
<td>295 (25.3%)</td>
<td>660 (27.0%)</td>
</tr>
<tr>
<td>T stage</td>
<td>T1</td>
<td>T2</td>
<td>T3 / T4</td>
</tr>
<tr>
<td></td>
<td>333 (26.1%)</td>
<td>610 (47.8%)</td>
<td>943 (38.5%)</td>
</tr>
<tr>
<td></td>
<td>387 (33.2%)</td>
<td>475 (40.7%)</td>
<td>862 (35.2%)</td>
</tr>
<tr>
<td></td>
<td>720 (29.5%)</td>
<td>1085 (44.4%)</td>
<td>1805 (73.7%)</td>
</tr>
<tr>
<td>N stage</td>
<td>N0</td>
<td>N1</td>
<td>N2 / N3</td>
</tr>
<tr>
<td></td>
<td>973 (76.3%)</td>
<td>223 (17.5%)</td>
<td>1196 (48.8%)</td>
</tr>
<tr>
<td></td>
<td>891 (76.3%)</td>
<td>141 (12.1%)</td>
<td>1032 (42.2%)</td>
</tr>
<tr>
<td></td>
<td>1864 (76.3%)</td>
<td>364 (14.9%)</td>
<td>2228 (88.1%)</td>
</tr>
<tr>
<td>M stage</td>
<td>M0</td>
<td>M1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1207 (94.7%)</td>
<td>8 (0.6%)</td>
<td>1215 (95.3%)</td>
</tr>
<tr>
<td></td>
<td>1039 (89.0%)</td>
<td>1 (0.1%)</td>
<td>1040 (89.1%)</td>
</tr>
<tr>
<td></td>
<td>2246 (92.0%)</td>
<td>9 (0.4%)</td>
<td>2255 (92.4%)</td>
</tr>
<tr>
<td>Tumor size</td>
<td>Mean (range)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>36.0 (0.8-110.0)</td>
<td>33.2 (1.5-110.0)</td>
<td>34.6 (0.8-110.0)</td>
</tr>
<tr>
<td>ECOG status</td>
<td>0</td>
<td>1</td>
<td>2 / 3 / 4</td>
</tr>
<tr>
<td></td>
<td>623 (48.9%)</td>
<td>558 (43.8%)</td>
<td>1181 (48.1%)</td>
</tr>
<tr>
<td></td>
<td>765 (65.6%)</td>
<td>362 (31.0%)</td>
<td>1127 (46.2%)</td>
</tr>
<tr>
<td></td>
<td>1388 (56.8%)</td>
<td>920 (37.7%)</td>
<td>2308 (92.5%)</td>
</tr>
<tr>
<td>Tumor Differentiation</td>
<td>Poorly</td>
<td>Moderately</td>
<td>Well</td>
</tr>
<tr>
<td></td>
<td>98 (7.7%)</td>
<td>140 (11.0%)</td>
<td>238 (9.7%)</td>
</tr>
<tr>
<td></td>
<td>79 (6.8%)</td>
<td>171 (14.7%)</td>
<td>250 (10.4%)</td>
</tr>
<tr>
<td></td>
<td>177 (7.2%)</td>
<td>311 (12.7%)</td>
<td>488 (19.2%)</td>
</tr>
<tr>
<td></td>
<td>196 (8.0%)</td>
<td>196 (8.0%)</td>
<td></td>
</tr>
</tbody>
</table>

Related factors[^10]. For example, (author?)[^20] have studied the associations between age, comorbidity, and other patient factors and treatment of postoperative NSCLC recurrence where they used a logistic regression to predict the postoperative prescribed treatment type – active or palliative. Supervised methods is able to operate on different types of data such as images, sequence and structured tabular data, therefore, they were utilized in different studies to consume heterogeneous data types. For example, they have been used for survival prediction of NSCLC patients from a combination of clinical and micro-array data[^15]. They were also used to model PET/CT images in order to predict early stage NSCLC patients survival[^21].

### Patient Cohort Analysis

Our study uses electronic health records (EHRs) of patients which are collected and stored by Spanish Lung Cancer Group (SLCG) which is a multi-disciplinary group focused on finding better treatments for lung cancer. In the following, we discuss the patient cohort examined in our study and the features and characteristics extracted from the patient EHRs.
• **Patient cohort.** Our study is conducted on a cohort of 2442 early-stage (stage I or II) NSCLC patients where 1275 (52.2%) of these patients had tumor recurrence after successful treatment. The mean age of the patients in our dataset is 65.6 years with little difference between recurrence (65.9) and disease-free survival (65.4) patients. Males represent around 74% of the patient cohort while 85% of this cohort are current or previous smokers. All the examined patients are diagnosed with stage I and stage II NSCLC where patients are split between the two stages with a 54.6% for stage I and 45.4% for stage II. In terms of the TNM staging system, the vast majority of patients have been diagnosed with M stage and N stage of 0 (92% and 76.3% respectively) while the T stage diagnoses has a more even split. The mean tumor size was 34.6mm, where patients who suffered a recurrence had a slightly larger mean tumor size (36.0mm) compared to disease-free survival surviving patients who have a mean tumor size of 33.2mm with a similar range for both. The ECOG performance status of the patient cohort is mostly divided between the statuses 0 and 1 with approximately 57% and 38% respectively while other ECOG statuses 2, 3 and 4 are associated with approximately 5% of the patients only. Table 1 provides a more detailed view of the characteristics of our patient cohort in relation to patient tumor recurrence and disease free survival.

• **Feature analysis.** The electronic health records of our patient cohort contain a different set of patient information and features such as the patient’s demographics, diagnosis, bio-markers, treatments and follow up information. Patient history includes gender, smoking habits, familial cancer history and comorbidities. Diagnosis features detail information on the tumour classification, histology, etc, at time of diagnosis along with the symptoms the patient suffered. The bio-marker features record the results of any bio-marker analysis performed on the patient during the course of their treatments. The treatment records of a patient include the details of any chemotherapy, radiotherapy or surgical procedures the patient underwent during their treatment. Chemotherapy features include the drugs used (including any maintenance drugs), the start and end dates of the treatment and the patients response to the chemotherapy. Radiotherapy features include the area radiated, the dose and the fractioning. Surgery features contain the type of surgical procedure, resection degree, the response and the post-surgical TNM staging values.

• **Patient labelling.** In order to label patients as having relapsed (i.e., positive cases for training the machine learning models) we rely on 2 feature groups in the SLCG dataset. The primary source for labelling comes from the patients’ progression records. If a patient has a progression record with status "Progression" or "Relapse" they are labelled positive. A secondary source for labelling comes from the patients’ follow-up records. If a patient has a follow-up status of "Alive with disease" or a follow-up status of "Dead" with cause of death "Lung cancer" they are also labelled as positive. All other patients are considered negative examples for the purposes of training our models.

• **Patient consent.** The patient data used in this study was collected under the provisions of the Law 14/2007 on Biomedical Research at all times along with the confidentiality of the data of patients according to the requirements of the law 15/9 [54] on Protection of Personal Data and Implications of EU General Data Protection Regulation 2016/679 (GDPR), applicable from 25 May 2018, Directive 95/46/EC.

The data collected was identified with a code. Access to patients’ personal information was restricted to the study doctors/investigators, the ethical clinical research committee of Puerta de Hierro-Majadahonda University Hospital and specialized authorized personnel to verify the study data and procedures, always keeping the confidentiality of the latter in accordance with current legislation. Only clinical data regarding the disease were used for subsequent analysis by the technical partners of the consortium, no personal data that could reveal the patients’ identity was used. The signature of the informed consent for entry into the study is revocable since the patients can withdraw their consent at any time so that we stop using their personal data or their own samples for this research; this revocation can be done simply by notifying the investigator. The entry in this study was totally voluntary so in case of revocation or refusal to enter the study the patient would continue to be clinically treated by their doctor whether or not participating in the study.

**Methods**

In this section we discuss the data preparation and preprocessing step, analysis of the features included in the investigated patients EHRs and the learning models that we use in our study.

• **Data preprocessing.** We have applied a set of data preprocessing operations on the patients’ EHRs to be consumable by our learning models and to decrease feature sparsity. We have fixed some corrupted date values where the date contains an invalid day or month value by replacing these values with a a placeholder value 1. We have also filled some
empty boolean fields with expected default values. For example, empty comorbidities annotations were assumed absent. This has been executed carefully with manual expert support to enrich the patient features extracted from their EHRs.

We have then formulated the patient records in a tabular dataset form, where we convert each patient record into a row of features. In this step, we have one included numerical and categorical feature values and we have excluded free text annotations and date values. We have attempted to preserve the order events in the patient row values by using different indexed columns for the same event which are indexed based on their order.

• **Data censorship.** In our study, we use the patient records to the full of their length to decide whether the patient had a tumour recurrence where we use all the available follow-up data points to detect recurrence. On the other hand, when build the patient features, we censor the patient data features using two criteria defined by experts to ensure that the learning models are only trained on features which are extracted from events successful treatment. First, if the patient had chemotherapy and radiotherapy treatments and did not have a surgery, we only consider the patient’s features prior to the first successful chemotherapy (the chemotherapy which results in absence of tumor). Secondly, when the patient have a successful surgery, we only consider features prior to this surgery and including the surgery detail.

• **Learning models.** In this study, we use a set of four supervised machine learning models: Random Forest (RF), Multi-Layer Perceptron (MLP), Support Vector Machine (SVM) and Logistic Regression (LR). In the following, we provide a short description of each of these models and how they work:

  * Random Forests are supervised machine learning models which uses a large number of small decision trees called estimators to produce predictions, where aggregates the prediction of these estimators to produce a more robust overall prediction. Random Forests can provide explanations for their predictions based on its decision trees (estimators) feature weights.

  * Multi-Layer Perceptrons are artificial neural networks composed of multiple neural layers, where each layer is composed by an affine transformation followed by an element-wise non-linear operation. In our implementation, MLPs use stochastic gradient descent to optimize the log-loss on the training data.

  * Support Vector Machines are supervised machine learning models which performs binary classification by learning a linear decision boundary by optimizing a margin-based loss function.

  * Logistic Regression is a machine learning model which assumes a linear relation between the input features and the output predictions. The predictions of the logistic regression model are easily interpreted in terms of the coefficients corresponding to each of the input features contributing to the final model prediction.

• **Model training pipeline.** We trained the previously mentioned models on the patient EHRs using a multiphase-procedure as illustrated in Fig. 1 where we perform the data preprocessing and censorship as previously discussed. We then train the models using a k-fold cross-validation strategy, where models’ hyperparameter tuning process is executed using the hyperparameters optimization framework optuna across at least 10 trials per model class. Further discussion on the model training and evaluation process is included in the Results section.
Table 2: A Comparison between four supervised learning models on 5-fold cross validation to predict recurrence in NSCLC patients as a binary classification task.

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F1 Score</th>
<th>Avg. Precision</th>
<th>AUC-ROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Forest</td>
<td>0.69 (0.032)</td>
<td>0.71 (0.028)</td>
<td>0.74 (0.059)</td>
<td>0.72 (0.034)</td>
<td>0.67 (0.023)</td>
<td>0.68 (0.031)</td>
</tr>
<tr>
<td>SVM</td>
<td>0.66 (0.015)</td>
<td>0.7 (0.015)</td>
<td>0.69 (0.027)</td>
<td>0.69 (0.016)</td>
<td>0.65 (0.011)</td>
<td>0.66 (0.015)</td>
</tr>
<tr>
<td>Neural Net. (MLP)</td>
<td>0.64 (0.019)</td>
<td>0.67 (0.019)</td>
<td>0.7 (0.026)</td>
<td>0.69 (0.018)</td>
<td>0.64 (0.014)</td>
<td>0.64 (0.019)</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>0.65 (0.019)</td>
<td>0.68 (0.018)</td>
<td>0.68 (0.026)</td>
<td>0.68 (0.019)</td>
<td>0.64 (0.014)</td>
<td>0.64 (0.019)</td>
</tr>
</tbody>
</table>

Figure 2: (a) Receiver operating curve (ROC) and (b) Precision-recall (PR) curve of the best performing models across 10 trials.

Results

In this section, we discuss the setup of our experiments where we describe the details of our evaluation protocol, metrics and the outcomes of our evaluation.

• Evaluation protocol. We cast the problem of assessing risk of tumour recurrence for a patient as a binary classification task which answers the question: whether a patient will have a recurrence or not. We execute our task on a cohort of 2242 patients where we use a 5-fold cross validation strategy. The 5-fold cross validation strategy is an evaluation technique where we split the patient cohort into 5 different random splits and evaluated our learning model on the five different splits independently. In each split evaluation, we use the other splits as training data and we evaluated our model on the specified splits. This technique produces a generalized view of the model predictive accuracy on the whole patient cohort with no data leakage between the training and testing cohorts.

• Metrics. We report the evaluation results of our examined models using a range of conventional binary classification metrics such as accuracy, precision, recall and the F1 score. We also use ranking metrics such as the average precision (i.e. the area under the precision-recall curve) and the are under the ROC curve (AUC-ROC).

• Model hyperparameter tuning. Our experiments were focused on four models as described in the methods section: Random Forest (RF), Logistic Regression (LR), Multi Layer Perceptron (MLP), and Support Vector Machine (SVM). The hyperparameters for these models were chosen using a hyperparameter search procedure executed using the hyperparameter optimization framework optuna. For each model, we execute a set of 10 trials to find the optimal hyperparameter configuration where each trial corresponding to a single hyperparameter configuration chosen from a predefined search space. We then choose hyperparameters corresponding to the best performing configuration as the
Figure 3: One of the best performing Logistic Regression model’s coefficients. As a linear model, LR is interpretable through its coefficients that are tuned to give weights to the input features. Top chart presents features yielded by model as relevant (coefficients > 0) which contributes the most to recurrence prediction, while the bottom one non-relevant for prediction (coefs <= 0) which contributes the most to predicting disease-free survival.

model best hyperparameters for each of the examined models.

• Evaluation results. Table 2 shows the results of our computational experimental evaluation of the four examined models using a 5-fold cross validation evaluation strategy. The results show that the random forest model achieved the best results in terms of all the used evaluation metrics where it achieves 0.69, 0.71, 0.74, 0.72, 0.67 and 0.68 scores in terms of accuracy, precision, recall, F1 score, average precision and the are under the ROC curve metrics respectively. However, the results also show that the margin between the metric scores of the random forest model and the other examined models in not significantly large, where the standard deviation of the metric scores of all the models is 0.02, 0.02, 0.03, 0.02, 0.01 and 0.02 in respect to the accuracy, precision, recall, F1 score, average precision and the are under the ROC curve metrics scores respectively. The results also show that all the examined model have consistent results through the different 5-fold cross validation runs where the reported standard deviation of the their reported metrics is always less than 0.05.

In the Fig 2 we demonstrate plots of the ROC and precision recall curves of the examined models. The ROC curve shows that all the examined models a better predictive accuracy than random the random baseline with marginal differences between the performance of all the models. On the other hand, the precision-recall curve demonstrate some disparities between the model performance especially in terms of the models precision in relation to changes of the model’s recall. For example, the logistic regression and SVM model suffer from a significant decrease of precision when the model’s recall is less than 0.2 in comparison to other models.

• Feature importance analysis. In Fig 3 we provide an example of model feature prioritization, where demonstrate the highest and lowest coefficient values learnt by the logistic regression model and their corresponding features. These coefficients represent the weight of contribution of their corresponding features towards the recurrence/survival predictions. The top part of figure demonstrates the features with the highest coefficient which contributes the most to
predicting that a patient would have a recurrence. For example, the highest contributing feature to predicting recurrence is have a stable disease at the end of the last successful chemotherapy treatment. Other major contributing features include smoking, previous cancers, comorbidities such as hepatitis and COB disease, and large tumor size at diagnosis represented by the T stage. The bottom part of Fig. 3, however, demonstrates the features with the lowest coefficient (negative values) which contributes the most to predicting that a patient would have a disease-free survival. The figure shows that the most significant features in this category are outcomes of the last successful surgery e.g. TNM stages and resection grade, and the final response to last successful chemotherapy treatments.

Discussion

The main goal of this work has been to assess the suitability of off-the-shelf machine learning models to the task of predicting relapse in early stage non-small cell lung cancer patients. To that end, we have used data on a comprehensive patient cohort from the Medical Oncology Department of Hospital Universitario Puerta de Hierro Majadahonda in Madrid, Spain. While the standard methods used in this context (such as models based on Cox proportional hazard analysis) were reported to produce competitive results, they do have their limitations.

First, they use purely statistical, local features based on frequency for the predictions, and thus they cannot, by definition, use complex, long-range relationships between the particular data elements. This is an aspect some readily-available binary classification machine learning models such as neural networks may address better, as they can naturally model non-linear boundaries between the classes of relapsing and non-relapsing patients.

The Cox models also cannot work with heterogeneous data (e.g. patient data together with biomedical database and publication data), even though such data are more likely to provide a truly comprehensive representation of the patients and their disease. This is an aspect models based on extensible and semantically integrated machine learning data sets, which we plan to incorporate in the next stages of this work, can address more naturally than purely statistical models.

Furthermore, using predictive models based on machine learning allows for rapid prototyping and easy experimentation with a broad range of possibly complementary models, as shown in this preliminary work.

While a rigorous clinical validation of the model predictions is a part of our future work, our experiments have demonstrated the feasibility of the machine learning-based approach in the context of relapse prediction. All explored models significantly outperformed a naive baseline in standard machine learning metrics, and thus established a solid non-trivial stepping stone for further experiments in this area.

Perhaps the most important lesson learned was the crucial role of clinical data preprocessing with direct involvement of the oncologists, who were instrumental in defining the scope of features applicable for training the machine learning models. They were also essential for devising specific strategy for labelling positive and negative patient examples that is a must for optimal and clinically-relevant model performance. We believe the strategy is generally applicable across similar datasets and clinical needs. However, there is certainly lot of space for improvement in terms of data cleansing and conversion of qualitative values of patient features into quantitative data that are more suited for the machine learning models. This aspect would require a dedicated publication on its own and is a part of our future work.

Another critical aspect we intend to address in future research is the clinical validation of the predictive models. One would ideally like to compare the performance of the models with a human baseline, which is, however, possible only in a prospective study we are currently performing. In these settings, the models will be trained on retrospective data and tested on prospective patients, which will let us assess the capability of the models to predict high risk scores of patients who did indeed relapse, and low scores of patients who did not, in the follow-up period. This will allow for much more realistic assessment of the machine learning models, and overcome the main limitation of the presented preliminary work.

Funding

This paper is part of the CLARIFY project that has received funding from the European Union’s Horizon 2020 Research and Innovation Programme under grant agreement No. 875160. It has also been funded by the Insight Centre for Data Analytics at National University of Ireland Galway (supported by the Science Foundation Ireland grant 12/RC/2289_2).
References


Word Embedding and Clustering for Patient-Centered Redesign of Appointment Scheduling in Ambulatory Care Settings

Iman Mohammadi, MS, PhD1,5, Saeed Mehrabi, PhD, FAMIA2, Bryce Sutton, PhD3, Huanmei Wu, PhD4,5
1Avalere Health, An Inovalon Company, Washington, DC; 2Secure Exchange Solution, Rockville, MD; 3Formerly Avalere Health, An Inovalon Company, Washington, DC; 4Temple University College of Public Health, Philadelphia, PA; 5Formerly Indiana University, Indianapolis, IN

Abstract

Background. A key to a more efficient scheduling systems is to ensure appointments are designed to meet patient’s needs and to design and simplify appointment scheduling less prone to error. Electronic Health Records (EHR) consist of valuable information about patient characteristics and their healthcare needs. The aim of this study is to utilize information from structured and unstructured EHR data to redesign appointment scheduling in community health clinics. Methods. We used Global Vectors for Word Representation, a word embedding approach, on free text field “scheduler note” to cluster patients into groups based on similarities of reasons for appointment. We then redesigned an appointment scheduling template with new types and durations based on the clusters. We compared the current appointment scheduling system and our proposed system by predicting and evaluating clinic performance measures such as patient time spent in-clinic and number of additional patients to accommodate. Results. We collected 17,722 encounters of an urban community health clinic in 2014 including 102 unique types recorded in the EHR. Following data processing, word embedding implementation, and clustering, appointment types were grouped into 10 clusters. The proposed scheduling template could open space to see overall an additional 716 patients per year and decrease patient in-clinic time by 3.6 minutes on average (p-value<0.0001). Conclusions. We found word embedding, that is an NLP approach, can be used to extract information from schedulers notes for improving scheduling systems. Unsupervised machine learning approach can be applied to simplify appointment scheduling in CHCs. Patient-centered appointment scheduling can be achieved by simplifying and redesigning appointment types and durations that could improve performance measures, such as increasing availability of time and patient satisfaction.

Keywords
Word Embedding, Clustering, Electronic Health Records, Appointment Scheduling, Healthcare Processes

Introduction

Appointment scheduling in health care is complex. Dynamic patient’s medical, physiologic, and mental state can lead to uncertainty in patient flow1. In acute health care systems, triaging is applied to evaluate acuity and meet demands. However, in non-acute or outpatient settings, triage scheduling is not the most effective way of scheduling. Outpatient settings should consider various factors such as: the number of services, the number of providers, the patient arrival process, the number of appointments, service times, and provider punctuality to design their scheduling systems2, 3. Our project goal is to redesign appointment scheduling to meet the needs of patients. In our previous study to improve access to care for underserved populations, we partnered with seven Community Health Centers (CHCs) which are ambulatory care settings providing primary and mental care for underserved populations and are designed as safety nets for these populations4, 5. We focused on discovering population needs, barriers to accessing healthcare, and strategies to reduce access barriers. We utilized Electronic Health Records (EHRs) covering a wide range of different information, consisting of both unstructured narrative text as well as structured data. We found redesigning appointment scheduling is one intervention that could potentially improve access to care. In this paper, we propose a patient-centered appointment template redesign by leveraging EHR data from our partner clinics.

Many appointment scheduling methods have been developed to address issues such as demand uncertainty, urgent care, and no-shows aiming at improving access to care and clinic service quality6. Other studies aimed at open access scheduling which allows patients to see a provider on the same day of requesting an appointment7. One general conclusion of these studies is that simplifying appointment types is an essential principle in implementing open access scheduling8,9-10. Simpler appointment types can reduce complexity in scheduling, leading to less error and better access to care11. A well redesigned scheduling scheme based on patient characteristics can improve utilization of medical
resources\textsuperscript{12}. In these studies, decreasing the number of appointment types or simplifying the appointment types in scheduling systems was recognized as a key step towards successful implementation. However, in these studies, there are no in-depth discussions on the appropriate ways to simplify appointment types.

Through our introductory literature review, we found that previous work in redesigning appointment scheduling did not propose patient-centered appointment structure for optimizing scheduling systems. Few studies focused on improving appointment scheduling based on patient characteristics, but these proposed scheduling systems were designed to accommodate health care settings like emergency departments, radiology departments, and inpatient settings rather than helping community health centers or outpatient settings. In this paper, we utilize real-world encounter data in community health clinics to identify appointment needs of populations they serve. We discuss how to leverage patient’s encounter data including reasons for seeking care to construct patient-centered appointment scheduling. We utilize a natural language processing approach, “word embedding”, to extract significant information from patient records. We use extracted information to cluster patients into groups based on similarity of their reasons for seeking health. The patient clusters are critical input in redesigning appointment types and durations that are simpler and more efficient without adding additional burden on clinics. Clinic managers and other stakeholders are encouraged to use the findings of this study to restructure their health care systems. This approach can also be a roadmap for developing automated appointment scheduling tools for ambulatory care settings.

**Methods**

Natural Language Processing (NLP) has been widely used to enable computers to understand free text and use the information derived from free texts\textsuperscript{13}. NLP includes wide range of computational techniques used by machines to comprehend human-like language processing\textsuperscript{14}. Word embedding is one of the featured learning techniques in NLP where words, phrases, or sentences are mapped to vectors of numbers\textsuperscript{15}. Word embedding can be used to derive semantic relationships between words using deep learning algorithms\textsuperscript{16}. Many studies in areas, such as sentiment analysis, information retrieval, and information extractions have applied word embedding\textsuperscript{16}. The source of free text data in our project is “patient reason for seeking health”. This field is entered by schedulers into the EHR systems. The objective of this study is to utilize word embedding to extract information from reasons for appointments, and then aggregate the similar reasons into single concepts. Those concepts are used to create new appointment types and durations. Figure 1 shows analysis engines used for redesigning appointment scheduling templates.

**Figure 1.** Analysis engines used for redesigning appointment scheduling templates. Unstructured data were used in Notes, MedTagger, WORD2VEC, and Validation steps. Structured data were used in Clustering, Validation, and Redesign steps.

**Data collection and preprocesing (figure 1, steps: Notes, MedTagger):** We collected EHR data from an urban community health clinic which included patient, visit, and provider characteristics. The field, “schedulers’ notes”, was the main data points to extract information. A scheduler note documents a patient’s reason for seeking care at the clinic. For example, when patient calls the clinic and asks for an appointment, the scheduler enters the patient explanation into the EHR system. We used schedulers’ notes to cluster patients based on the similarity of reasons for seeking healthcare. Schedulers’ notes are free text fields with many abbreviations; therefore, any attempt to extract information should include resolving issue of abbreviations. MedTagger, a library developed by Mayo Clinic, contains a suite of programs indexing based on dictionaries\textsuperscript{17}. We used the MedTagger\textsuperscript{18} dictionary list to expand the abbreviations to their full forms, for example “DM” is transformed to “diabetes mellitus”.

**Text mining (figure 1, step: WORD2VEC):** There are various word embedding models that map words to vectors (word2vec) of real numbers. These methods are generally categorized into two methods of matrix factorizations and shallow window-based models. Matrix factorization methods capture the statistical information about the corpus. Approaches, such as latent semantic analysis (LSA)\textsuperscript{19} capturing the term document frequencies or Hyperspace Analogue to Language (HAL)\textsuperscript{20} capturing the term-term frequency are two examples of matrix factorization methods.
The problems with these methods are that the most frequently used words contribute a disproportionate amount to the similarity measure, for instance co-occurrence with words such as “the” or “a” has a large effect on the similarity measure despite a lack of semantic relatedness. The skip-gram and continuous bag-of-words (CBOW) models are two of the most widely used word2vec approaches that use neural network structures in learning word representations. We used the Global Vectors for Word Representation (GloVe) method to represent each text column with their real-valued vectors. The GloVe model addresses the shortcomings of the earlier models. GloVe captures the benefits of count data while simultaneously capturing the meaningful linear substructures prevalent in modern log-bilinear prediction-based methods. We used the GloVe pre-trained vectors on a 6 billion token corpus from 2014 Wikipedia to construct a 50-dimensional vector for every word in the text that appeared in the pre-trained model. We normalized vector for words without representation in the pre-trained vector model; more specifically, we assigned the average of words appeared in the body of our sample to words without representation. We then averaged all the vectors for the words in the sentence to calculate the final representation of each sentence. For each patient encounter in the EHR, there is a scheduler note in free text format. We ran the word2vec algorithm on each encounter note. Each encounter was converted to a row with 50 columns representing the 50-dimensional vector that is derived from the note.

**Clustering (figure 1, step: clustering, validation):** The data were then fed into an Agglomerative clustering algorithm. Agglomerative clustering is a bottom-up hierarchical clustering approach by merging pair(s) of clusters in which clusters generated in earlier step might be nested within the ones generated later. This approach does not necessarily neglect the small clusters; hence, it is useful for the discovery of the smaller groups. To find the optimal number of clusters, we started with 2 clusters and stepwise increased the number of clusters to 20. In each run, we compared the results of clustering by analyzing the profile within each cluster. Attributes such as age, gender, and provider specialty were used to objectively validate the appropriate number of clusters. We also evaluated the clusters by reading 100 notes per cluster to see whether clustered notes are aligned with human judgment. We found the optimal number of clusters is between 10 to 12. We chose 10 as our final number of clusters for this study.

**Appointment type and duration redesign (figure 1, step: redesign):** In this step, the goal is to optimally give new appointment duration to each of the new 10 appointment types, i.e. the 10 clusters. We assumed that the clinic capacity and demand do not change. We investigated how standalone simplification of appointment types and durations could potentially impact performance measures, such as number of patients seen per year and patient satisfaction defined as patient time spent in-clinic. Patient time spent in-clinic is the difference between patient arrival and departure times and includes the sum of waiting time to see the provider, time with the provider, and time spent for check-out and payments. Proposed appointment durations were calculated based on the capacity that the clinic must accommodate daily. The sum of provider hours allocated to see patients per day was defined as daily clinic capacity. For example, if the clinic had two providers on a given day who each allocates 4 hours to see patients, the total capacity of the clinic on that day is 8 hours (i.e. 480 minutes), or 240 minutes per provider.

We used the distribution of current appointment durations per cluster to determine the most effective appointment durations for each cluster. For each cluster, we started by assigning the median current appointment durations to proposed durations (figure 2), then we increased it step by step by 1 percentile to the maximum. We then used the capacity and demand of the clinic to calculate performance measures in each step.

**Performance measures in the proposed appointment system:** Performance measures were:

1) Number of patients seen per year. The difference between current durations and proposed durations was calculated as time available to see more patients. We then calculated the number of additional patients that can be seen in the proposed system by dividing the time available to see more patients by new appointment durations per cluster. To normalize this measure, we calculated the number of additional patients that the clinic can see in the proposed scheduling system by year.

2) Provider time with patient. It was measured as total appointment duration (minute(s)) per provider.

3) Predicted patient time spent in-clinic. The time patients spent in-clinic, which includes in-clinic waiting time plus time spent seeing the provider, was calculated as the difference between arrival time and departure time recorded in the EHR data. We used current appointment durations, arrival time (AM vs PM), gender, provider specialty, number of provider(s) available in the day of appointment, day of week, and patient age as independent variables and in-clinic time as the dependent variable to develop a multivariate linear regression model. We used the regression model to predict time spent in clinic using the proposed appointment durations. Pairwise t-tests were calculated to test for significant differences between current and proposed systems.
We fit the current daily demand to the current daily capacity using the proposed types and durations. We found the most effective duration for each cluster by maximizing number of patients per year and provider time with patient while minimizing overall patient time spent in-clinic (figure 4).

Results

We collected 17,722 encounters of an urban community health clinic in 2014. The dataset included deidentified patient ID, day and time of encounter, patients’ arrival and departure times, age, gender, provider ID and specialty, appointment type (102 unique types recorded in the EHR), and appointment notes (or schedulers’ notes). The dataset included 7,061 unique patients in 2014. Following data processing, NLP implementation, and clustering, appointment types were grouped into 10 clusters using patients’ needs in the current scheduling systems (shown in table 1).

Table 1 and figure 2 were used to determine the most accurate number of clusters. Our proposed scheduling system has 10 types of appointments (noted as clusters). Table 1 shows examples of free texts that were aggregated into one concept. Cluster 1 seems to be appointments that are assigned to patients with complex issues. Cluster 2 represents acute female complications or patients with behavioral health needs. Cluster 3, which is the largest in terms of number of reasons, consists of acute care encounters that need to be scheduled as soon as possible. Clusters 6 and 7 are assigned to patients with chronic pain problems and other chronic problems. Clusters 8 and 9 are predominantly for pregnant, reproductive health and other female complications. Cluster 10 is for wellness and other childcare patients.

Figure 2 illustrates the distribution of current appointment durations per each cluster. Appointment durations typically range from 10 to 60 minutes. Cluster 1 has the highest durations, and this is aligned with the visit reasons shown in table 1, because it is given to complex patients. Cluster 10 has the lowest durations as it is given to well childcare.

Table 1. Cluster profiles and examples of reasons grouped into clusters.

<table>
<thead>
<tr>
<th>Clusters</th>
<th>Number of appointment types in current scheduling</th>
<th>Average appointment duration in current scheduling (min)</th>
<th>Number of unique reasons for visit</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6</td>
<td>20.0</td>
<td>203</td>
<td>knot on left breast is more tender and now hot touch not hot today has tried ibuprofen and tylenol f</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>colpo R/S due to + Trich in pap test and pt did not come in to R/O via urine Needs Urine Testing</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>hx of BV, vap, hx of, urine concentrated, but not now, burning on urination</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Bipolar, Anxiety med f/u.</td>
</tr>
<tr>
<td>2</td>
<td>27</td>
<td>17.6</td>
<td>859</td>
<td>lightheaded, vomiting, intermittent umbilical pain,</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>cough, congestion, runny nose, tired, decreased appetite</td>
</tr>
<tr>
<td>3</td>
<td>37</td>
<td>17.5</td>
<td>4646</td>
<td>birth control consult, here with involved mother has tried depo last IM 12/2013, reports not happy w</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>asthma check mother concern speech not clear –history of father having speech problems child</td>
</tr>
<tr>
<td>4</td>
<td>27</td>
<td>17.6</td>
<td>494</td>
<td>fu multiple ED visits for abdominal pain, N&amp;V - MCARE</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Postpartum, del 6/19; never had PP visit, wants depo</td>
</tr>
<tr>
<td>5</td>
<td>37</td>
<td>18.1</td>
<td>2619</td>
<td>low back pain, pain in legs Has been taking wife’s medications for pain</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>stomach and chest pain (pt wanted to wait until this day for appt)</td>
</tr>
<tr>
<td>6</td>
<td>30</td>
<td>16.8</td>
<td>2078</td>
<td>wants to get off work due to side effects of medication “other problems”</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>meds/gallstones-upper mid-abd.pain since Sat. -P/S 10 @times needs meds for bipolar</td>
</tr>
<tr>
<td>7</td>
<td>34</td>
<td>17.8</td>
<td>1589</td>
<td>New OB HX @10:15 nob packet given and instructed on verbal consent for uds and hiv declines mfn ref</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>ROB 37 wks wants cx checked, increase in contractions and increase in pressure</td>
</tr>
<tr>
<td>8</td>
<td>24</td>
<td>19.9</td>
<td>317</td>
<td>pregnancy symptoms, no period x10wks/ neg upt 01/14/14 trying to conceive x 5 years nausea, irritabl</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>F/u labs/pelvic pain pelvic pain x 2 week c/o clear vaginal discharge +odor-itch</td>
</tr>
<tr>
<td>9</td>
<td>31</td>
<td>18.9</td>
<td>728</td>
<td>9 month wcc cough, cold sx’s “switching” episodes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Well child 12 mos Commercial Insurance Vaccines UTD (CHIRPS Printed)</td>
</tr>
</tbody>
</table>
The overall clinic patient gender distribution was 63% to 37% for females and males respectively. Table 2 provides a breakdown of age and gender profiles for each cluster. Cluster 1 represents younger patients from both genders. Cluster 10 shows that 95% of patients are younger than 13 years old, and it represents a pediatric population. Clusters 8 and 9 consist of predominantly female patients. Cluster 3, that was determined to be acute care based on table 1, represents all ages and genders. The gender and age profile of each cluster seems to be in agreement with examples of reasons for visits in table 1.

Table 2. Distributions of patient age and gender within each cluster.

<table>
<thead>
<tr>
<th>Appointment cluster</th>
<th>Age</th>
<th>Gender</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>17</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>26</td>
<td>61</td>
</tr>
<tr>
<td>2</td>
<td>27</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>21</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>59</td>
</tr>
<tr>
<td>3</td>
<td>22</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>19</td>
<td>59</td>
</tr>
<tr>
<td></td>
<td>22</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>26</td>
<td>61</td>
</tr>
<tr>
<td></td>
<td>28</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>27</td>
<td>65</td>
</tr>
<tr>
<td></td>
<td>19</td>
<td>13</td>
</tr>
<tr>
<td>10</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>13</td>
</tr>
</tbody>
</table>

Figure 3 shows percentages of appointments within a cluster that were scheduled with various provider specialties. Cluster 2 is a mix of behavioral health and all other specialties. Cluster 3 (acute care) patients were scheduled with all types of specialties. Cluster 10 patients are predominantly scheduled with pediatricians.
Figure 4 shows scheduling performance measures per several potential durations for new appointment types. Performance measures are the percentage reduction in average patient in-clinic time, ratio of patients seen in a new practice compared to the current practice, and ratio of provider time spent with patient compared to their capacity. For example, if we consider the value of the 65th percentile of all durations within a cluster to the new appointment duration for that cluster, we would see an 11% increase in patient time in the clinic, a 35% increase in number of patients accommodated, and a ~30% decrease in provider time with patients. The results in figure 4 include iterations from the 65th to 80th percentiles. We did not see changes outside this range, so they are not included in the figure. We chose the 75th percentile duration of appointments within each cluster as the new proposed appointment duration, because it can reduce average patient in-clinic time by 10%, increasing the overall number of patients to be seen by 9%, without significantly affecting provider time spent with patient.

Figure 4. Performance measures by iterations. Nth iteration means assigning the Nth percentile of appointment durations within a cluster to the cluster.

Table 3 shows comparisons of the current scheduling system and the proposed scheduling system. Average appointment duration in the current scheduling system is the average of current durations by cluster. Averages of appointment duration in proposed (i.e. 75th percentile duration of appointments within each cluster) scheduling system are higher for clusters 1, 2, 4, 7, 8, 9, and 10, and lower for clusters 3, 5, and 6 compared to the average(s) of current durations (p-value<0.0001). The time patients spent in-clinic per visit is calculated based on the EHR patient’s arrival and departure times. Predicted time spent in-clinic was calculated using a linear regression model trained using the current scheduling. Table 3 shows the proposed scheduling system could open space to see overall an additional 716 patients per year, which is about 10 percent more patients. Figure 5 shows distributions of patient time spent in-clinic per visit. Our results suggest that the new scheduling systems and appointment duration could decrease patient in-clinic time by 3.6 minutes on average (p-value<0.0001).

<table>
<thead>
<tr>
<th>Appointment cluster</th>
<th>Average appointment duration in current scheduling system (minutes)</th>
<th>Average appointment duration in proposed scheduling system (minutes)</th>
<th>Average time spent in clinic in current scheduling system (minutes)</th>
<th>Average predicted time spent in clinic in proposed scheduling system (minutes)</th>
<th>Number of additional/less patients clinic can see in the proposed scheduling system (patient/year)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20.0</td>
<td>30</td>
<td>50.8</td>
<td>67.1</td>
<td>-4</td>
</tr>
<tr>
<td>2</td>
<td>17.6</td>
<td>20</td>
<td>63.0</td>
<td>61.5</td>
<td>-21</td>
</tr>
<tr>
<td>3</td>
<td>17.5</td>
<td>15</td>
<td>66.0</td>
<td>57.7</td>
<td>649</td>
</tr>
<tr>
<td>4</td>
<td>17.6</td>
<td>20</td>
<td>64.3</td>
<td>61.0</td>
<td>-58</td>
</tr>
<tr>
<td>5</td>
<td>18.1</td>
<td>15</td>
<td>68.0</td>
<td>57.4</td>
<td>368</td>
</tr>
<tr>
<td>6</td>
<td>16.8</td>
<td>15</td>
<td>64.7</td>
<td>57.6</td>
<td>115</td>
</tr>
<tr>
<td>7</td>
<td>17.8</td>
<td>20</td>
<td>66.9</td>
<td>60.9</td>
<td>-87</td>
</tr>
<tr>
<td>8</td>
<td>19.9</td>
<td>30</td>
<td>70.3</td>
<td>68.3</td>
<td>-61</td>
</tr>
<tr>
<td>9</td>
<td>18.9</td>
<td>20</td>
<td>67.5</td>
<td>61.7</td>
<td>-138</td>
</tr>
<tr>
<td>10</td>
<td>17.1</td>
<td>20</td>
<td>65.4</td>
<td>58.4</td>
<td>-47</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>716 (10%)</td>
</tr>
</tbody>
</table>
Discussions

We studied the possibility of using patients’ reasons for seeking health along with patient, visit, and provider characteristics to design new appointment types and durations for community health centers. Our study has three major findings. First, word embedding, that is an NLP approach, can be used to extract information from schedulers notes for improving scheduling systems. Second, unsupervised machine learning approach can be applied to simplify appointment scheduling in CHCs. Third, patient-centered appointment scheduling can be achieved by simplifying and redesigning appointment types and durations that could improve performance measures, such as increasing availability of time and patient satisfaction.

In this work, we expanded utilization of word embedding trained models by applying it on scheduler notes in primary care settings. We found word embedding trained on EHR scheduler notes using MedTagger, and GloVe can capture semantics of medical terms, and the results are aligned with human judgment (shown in table 1).

The Institute of Medicine defines health care quality as "the degree to which health care services for individuals and populations increase the likelihood of desired health outcomes and are consistent with current professional knowledge." One of the domains of health care quality is efficiency. Our study found simplification of scheduling based on patient, provider, and clinic characteristics could improve efficiency. In this work, we designed a methodology to simplify appointment types and times because complex schedule templates could lead to mismatching patient problems to incorrect solutions. Simplifying appointment types and times is one of the requirements of transitioning from traditional access models to advanced access models. The approaches in this study could simplify appointment scheduling to match daily supply and demand. We found simplifying scheduling templates could improve overall clinic performance, such as improving provider productivity, decreasing patient in-clinic waiting time, and improving clinic accommodations. Our methodology is significant because improved overall performance could be achieved without additional supply, more resources, or extended hours.

Patient-centeredness is another domain of health care quality that is achieved by meeting patient needs and preferences. In this study, we designed an infrastructure for patient specific resource allocation. Patients with different reasons for seeking health, age, and gender have different resource requirement. Our proposed appointment scheduling template clusters patients into classes based on reasons for seeking health. Timeliness and patient satisfaction are other aspects of a good health care delivery system. Our study found that simplified scheduling can reduce in-clinic time that could consequently lead to improved timeliness and satisfaction.

Our study has few limitations. First, our patient encounter data lacked clinical information such as diagnoses, procedures, lab results, and clinicians’ notes. In any future work these features can also be used to design stronger patient specific resource distribution. Another limitation of this study was that our dataset did not include information about in-clinic patient journeys, such as step by step activities and timestamps from the moment that a patient checks in to departure of patients, and information about daily staffing of medical assistants and nurses. Those factors could be predictors of in-clinic waiting time. Another limitation of the methodology is computationally resource intensive.
nature of agglomerative clustering, especially when it comes to large data. Other clustering or unsupervised learning methods might be explored. One limitation of word embedding model such as GloVe is that words must be seen in the training data in order to have an embedding. There are various methods to deal with out of vocabulary (OOV) words such as subword embedding, <unk> replacement, random initialization, etc.

Future work in this area might focus on four objectives. First, expansion of abbreviations by utilizing more comprehensive dictionaries that would include less commonly used abbreviations. Second, other unsupervised clustering methods such as deep learning or reinforcement learning might be able to extract more relations between notes which would lead to more precise clusters. Third, researchers might use the findings of this study to either implement the algorithms in current EHR interfaces or design a new interface for a decision support system. Future research in this area could also evaluate the effectiveness of the proposed algorithms in real world clinical practice. Forth, EHRs contain a longitudinal data and information on previous patients’ encounters that can be considered to augment this redesign approach.

Potential Medical Applications. One of the steps of moving from traditional appointment scheduling to optimized open access scheduling is to simplify appointment types and times. Ambulatory care settings can leverage methodologies and findings of this paper to achieve optimized open access scheduling. Previous studies did not discuss the most appropriate ways to simplify appointment types. These previous studies mainly offered appointment types such as “new”, “established”, “acute”, and “postoperative” as decreased number of appointments. A key advantage of the methodology presented in this paper is that the simplification of the appointment template not only helps clinics implement advanced open access scheduling system, it is also patient-centered and patient specific. The proposed appointment scheduling templates are designed based on reasons patients are seeking health care. Another potential medical application of this study is to utilize the unsupervised machine learning approach presented in this paper to design automated appointment scheduling tools for healthcare settings. These tools can be in the form of online appointment scheduling or automated phone call scheduling. These potential tools ask patients why they need appointments and the system finds the most appropriate appointment type and time for the patient. Methodologies presented in this paper can also be applied on both scheduler and clinician notes to find care needs and gaps for patients and design interventions to close the gaps.

Conclusion

A key to a more efficient scheduling systems is to ensure appointments are designed to meet patients’ needs, and to design and simplify appointment scheduling which is less prone to error. In this paper, we presented approaches for redesigning appointment scheduling based on patient characteristics, needs, and desires. We used EHR data to investigate the relationship between patient characteristics and reasons for visit to help providers redesign healthcare systems that can meet the needs of patients. We applied word embedding and unsupervised machine learning methods to design more effective and efficient appointments in ambulatory care settings. We found that simplifying appointment types and times can help healthcare systems achieve improved access and patient satisfaction without adding additional resources.

References

Agile Implementation of Innovative End to End Technical Solutions for Respiratory Testing in the COVID-19 Pandemic

Tamara Moores Todd, MD\textsuperscript{1,2,}, Kathryn G Kuttler, PhD\textsuperscript{1,}, Diego Ize-Ludlow, MD\textsuperscript{1}

\textsuperscript{1}Intermountain Healthcare, Salt Lake City, Utah
\textsuperscript{2}University of Utah, Salt Lake City, Utah

Abstract

The COVID-19 pandemic required rapid implementation testing at large scale. We describe our approach to creating an efficient and highly scalable end-to-end testing and reporting workflow. We utilized a combination of consumer-driven workflows, a consistent ordering and data collection EMR user interface, an EMR independent entry point for affiliated providers, and rapidly evolving automation of screening, registration, patient identification, curbside specimen management, and prioritization logic. Our agile approach allowed us to transition from people-driven processes and gain efficiencies utilizing previously developed informatics infrastructure, new solutions and emerging consumer digital tools. We supported over 1,000,000 tests with >200 prioritization logic changes without requiring education or changing the provider ordering workflows, handled patient-driven screening, data collection, and ordering of up to 9 patients/minute while decreasing curbside specimen collection from more than 39 to less than 17 minutes per patient. The approach supported texting consumers and standards-based reporting to public health agencies.

Introduction

The development and implementation of efficient testing strategies for SARS CoV-2 has been key to the public health response and care of patients impacted during the COVID-19 Pandemic. This report describes our informatics approach to the rapid scaling of standardized guided screening and ordering processes, prioritization, specimen collections, lab processing, and results reporting. Intermountain Healthcare is an integrated not-for-profit health system with more than 41,000 employees who serve the needs of people across the intermountain west, primarily in Utah and Idaho. The system includes clinics, a medical group integrated by 2,400 physicians and advanced practice providers, affiliate networks comprising 3,800 physicians, hospitals, homecare, telehealth and health insurance plans. The organization provided COVID-19 testing services to affiliated providers, to public health agencies, to business entities, as well as to patients being cared for by the health system.

An ongoing challenge during the pandemic has been limited testing supplies, equipment, lab capacity, and the associated rapid evolution of testing criteria. National and local guidelines rapidly shifted who should be tested, tests to be used for which clinical scenarios; this was compounded with the supply chain challenges and local evolution of the equipment available and improving turnaround times of the different tests. This resulted in the need to apply strict but changing criteria to testing. We needed to be able to provide intuitive guidance to the shifting testing criteria in electronic and paper-based ordering.

With Intermountain Healthcare’s unique geography, there was a need to create a consistent ordering workflow which would facilitate specimen collection throughout the states of Utah and Idaho, allow high-throughput testing in our centralized laboratory while also leveraging regional and external labs when testing capacities varied due to supply chain shortfalls. In addition, as Intermountain Healthcare provides care for a wide variety of clinical scenarios, we needed to ensure that the appropriate prioritization was assigned to the appropriate patient scenario, so that a critically ill patient might have the most rapid turnaround time in order to qualify for therapeutic intervention, and also to maintain our community responsibility to provide rapid turnaround testing to a large volume of the population in the outpatient setting.

Due to the rapidly changing environment of testing supply availability and changing epidemiology, our team was initially tasked with creating a stable, front-facing workflow which would automatically flex ordering criteria, require and document the U.S. Department of Health and Human Services Ask On Order Entry (AOE) questions, calculate

872
specimen prioritization, expedite lab processing, and patient resulting at a volume and rapidity never previously experienced in healthcare.

Ordering

In the face of such a rapidly changing environment, we aimed to surface clinical decision support within the providers’ natural workflow. The recommendations for testing would then either allow the healthcare team to proceed with COVID-19 testing based on the most up to date Infectious Disease recommendations or defer testing and recommend other management. The decision support would allow us to have a stable ordering interface for providers, without depending on providers to re-learn the rapidly changing testing guidelines via other lagging asynchronous methods (email, PDFs, department meetings, etc).

Beyond hosting our ordering criteria, AOE questions and prioritization within our Electronic Medical Record (EMR), Intermountain Healthcare also services many affiliated providers who partner with Intermountain for collections and testing, but who don’t normally access our EMR. We needed to create an easy, reproducible method for affiliated providers to meet the required elements mentioned above.

Expanded Screening

As the urgency of the pandemic increased, it became clear that a provider visit was not required for basic screening for testing eligibility in non-critically ill patients. There was a desire to use unlicensed individuals (ex. Medical Assistants) for screening via a COVID-19 phone hotline. We were challenged with the question of devising a simple method of showing the ordering criteria and facilitating registration, insurance gathering, ordering, and scheduling, in a manner sustainable by a standing-order protocol. By achieving this workflow, we hypothesized we would have the ability to dramatically increase our offerings to our community patient populations without jeopardizing patient safety.

Specimen Collection

Due to the risk of healthcare worker infection from COVID-19 specimen collection, by virtue of a nasopharyngeal swab being an aerosol generating procedure, Intermountain strategically decided to host all outpatient collections at specific curbside testing locations. This allowed the process to maximize safety and optimize personal protective equipment resources. The curbside collection sites would both collect samples ordered by other healthcare teams, and also provide on-site screening for eligibility of testing for patient walk-ins. At the peak of testing, Intermountain Healthcare maintained 26 curbside collection locations throughout the states of Utah and Idaho. Any inefficiency in the process of on-site registration, insurance collection, ID verification and specimen collection would be a large rate-limiting part of the pandemic response.

Lab

In early March 2020, Intermountain Central Laboratory began COVID-19 testing with the ability to process and result approximately 300 tests/day. At that time, there were only a few labs in Utah able to process COVID-19 tests. Due to extremely limited supply and large demand, each lab had variable capacity to perform tests on a given day. Each lab had multiple different order codes, depending on type of COVID-19 test available based on location and testing instrument. In the first deployment of testing, if Intermountain Central Lab ran out of capacity on a given day, the lab personnel would need to manually cancel and reorder all tests with the external lab order code, which was highly inefficient. In order to maximize efficacy we needed to devise a system of rapidly flexing the specific test order code available to Intermountain ordering providers depending on the day of availability without changing the providers’ ordering workflow in a noticeable manner, thereby relieving the lab personnel of the need to manually cancel and reorder thousands of tests in a given day.

Reporting

As important as it is to provide interventional healthcare after infection in a nascent pandemic, facilitating a convergence with public health measures to prevent further community spread of infection remains essential to a comprehensive response. Therefore, we needed to support public health reporting on an extraordinarily broad and rapid scale, while not hindering our management of actively infected patients. While Intermountain was already a recognized participant in Electronic Case Reporting (eCR) several challenges still existed, which required nimble collaboration and innovation between Intermountain, state and federal data systems to ensure timely contact tracing and other essential public health measures.
Beyond reporting innumerable data regarding patient presentation and test results to public health entities, the burden of immediately reporting test results and counselling directly to the patient remains elder to all other measures. While our initial processes were quickly deployed people-driven results callbacks, the pivot to technology-based direct reporting of results to patients was urgent in our evolution.

**Methods**

**Ordering Process**

At the beginning of the pandemic SARS-CoV-2 by PCR tests (COVID-19 tests) were in short supply. It was important to preserve these valuable resources and prioritize COVID-19 tests for different patient populations.

In such a rapidly changing environment regarding testing criteria, epidemiology and recommendations, we made the strategic decision to host Clinical Decision Support within the normal ordering workflow. Selecting a COVID-19 testing order triggered a universal COVID-19 screening data intake form, which flexed the criteria logic face up, telling the healthcare team if the patient qualified for testing based on their answers on the intake form. If the patient met testing criteria, the healthcare team could proceed and order the test. By channeling all COVID-19 testing via this intake form, we were able to ensure continuity of practice across a wide geography and variable clinical environments, ensure data reporting was performed on each COVID-19 test, and leverage the various data points for internal and external review regarding trends within the pandemic on a real-time basis.

![COVID-19 Screening Form](image)

*Figure 1.* Universal COVID-19 screening data intake form. Caregiver input of discrete data were used in Decision Support algorithms, ensured continuity of practice, enabled agile changes during the pandemic, ensured data reporting on each COVID-19 test, and leveraged the various data points for internal and external review regarding trends within the pandemic on a real-time basis.
Furthermore, once we created this framework of ordering design, it was simple to quickly change the testing criteria on a compressed delivery timeline. Additionally, the change to ordering logic would be immediately reflected in active patient workflows. For illustration, initially the only criteria which would allow a patient to have a COVID-19 test based on CDC guidelines was presence of 1. Fever OR 2. Cough OR 3. Shortness of Breath. This rapidly expanded, and as Intermountain remained aligned with the State of Utah and Federal testing guidelines, we were able to leverage our technical tools to deliver and facilitate information to the end-user regarding the most up to date criteria in a stable and actionable workflow. Once the intake form was established, we were able to utilize the universal criteria as the basis for standing order protocols. Our ability to leverage unlicensed staff to screen patients for COVID-19 testing became easily facilitated because the entire organization was following a set list of criteria for patient testing. We created order sets including a standing order, and auto added the correct diagnosis code based on clinical testing indication provided within the intake form, to ensure appropriate data reporting, billing and coding.

For affiliate provider ordering, we created an affiliate website which displayed the ordering criteria currently accepted at Intermountain Healthcare. A lab requisition in the form of a fillable PDF was provided, with the option to e-mail or fax the form to a universal inbox. The unlicensed staff of the COVID-19 hotline would monitor the inbox and enter the orders into our EMR using the standard intake form, attributing the ordering physician as the affiliated physician. The patient would then present to one of our curbside collection sites for specimen collection.

Rules Based Ordering and Prioritization Logic

As previously described, beyond simply providing access to a large volume of testing, there remained a clinical imperative to balance our testing prioritization between our acute inpatient patient populations and our community health directive. To put it simply, a critically ill intubated ICU patient with potential for novel therapeutics was deemed to need a quicker result compared to an otherwise healthy individual in the community getting screened due to close contact with a COVID-19 case. As we investigated various options to calculate prioritization and its associated workflow, the options initially evaluated included 1. Manual printing of order form, with lab personnel calculating patient priority by hand; 2. Making the ordering providers note the patient priority as a part of ordering workflow; 3. Automated calculation of prioritization using Boolean-based logic, and printing prioritization directly on lab order label for lab personnel to read when scanning the specimen. After evaluating each option, we elected to choose option 3. This choice allowed us to flex testing instrument order code and prioritization in a dynamic fashion, while never changing the ordering interface for providers. Depending on the patient presentation and available data, we were able to change the patient prioritization algorithm on a daily or hourly basis depending on direction from lab and infectious disease leadership without any change in the clinical healthcare team’s workflow.

Laboratory Specimen Prioritization

Boolean logic-based prioritization algorithm was developed to flex the appropriate type of test and test prioritization after the intake form was completed. Clinical Informatics huddled with lab, Infectious Disease, and Clinical Operations leadership every day and decided on prioritization adjustments which were synchronously implemented. While this methodology was efficient for the frontline clinical workflow, it required significant work from the Clinical Decision Support team to flex the logic based on internal and external lab availability on a daily basis. In order to smooth this process, an external lab configuration webpage was developed as a user interface, which allowed the lab leadership to directly choose which prioritization levels would send to internal or external lab sites. By placing the prioritization algorithm order codes directly in the hands of the lab leadership, they were able to change a simple UI with one click. This modification would instantly change the order code and lab assignment placed by the decision support algorithm for all COVID-19 orders of a certain priority throughout the Intermountain EMR. This allowed our system to remain extraordinarily nimble while handling rapidly increasing volumes of collections and testing across a two-state geography.

In addition, in order to further develop lab efficiency, we facilitated the ability of the lab to directly scan the specimen sample’s (e.g. saliva) barcode directly into the Lab information system, rather than typing the number into the system. This simple fix was exponentially advantageous in increasing lab capacity and decreasing lab turn-around time.

Curbside Collections

As we streamlined the screening and ordering workflow, attention then shifted to the community collection process. In order to optimize the registration, identity (ID) verification, and collection process, we developed a novel workflow
leveraging EMR agnostic text messaging/emailing with a JavaScript app hosted in our EMR at the time of collection check-in.

For the automatic text message/email, we utilized PubSub\textsuperscript{2} to identify all future orders, regardless of ordering location (ex. pre-procedural vs. COVID-19 hotline vs affiliate ordering vs. Primary Care Provider (PCP) ordering), and then sent a text message and an e-mail to patients within 60 seconds of their order completing within our EMR. The message would contain a visit ID, a link to a scannable QR code, and also had an optional link for self-scheduling an arrival window to level-load patient volumes at the collection sites. The numerical visit ID/QR Code contained no protected health information and was a meaningless number if attempted to be scanned outside of Intermountain’s EMR.

Again, the simplicity of this solution was that every COVID-19 outpatient order generated within Intermountain Healthcare would generate a patient text message and email (as long as the patient had previously agreed to receive electronic communications), creating a consistent patient experience regardless of the specific use case for COVID-19 testing. We were able to split the patient cohorts of messages into their respective “reason for test” as indicated by our COVID-19 ordering intake form. Therefore, all the pre-procedural testing patients could have unique information sent to them specific to pre-procedural testing requirements vs. message content for general symptomatic patient education.

At the time of arrival to the curbside collection location, the patient presented their visit ID or QR Code to the healthcare team. By virtue of the patient having a text message/email containing this information, the curbside team knew with a glance that the patient registration, insurance, and ordering had already been completed. Prior to the implementation of this solution, the curbside collection team would re-verify the patient registration, insurance, and order placement on all patients, even if the patient stated they had a previously placed order by another healthcare team. Once the curbside healthcare team had confidence that the generation of a visit ID/QR Code reliably indicated complete patient registration, insurance capture and order placement, their efficiency markedly increased.

To maximize the impact of visit ID/QR code, we coupled this deployment with a de novo web module to facilitate patient check-in at the curbside collection locations. The website asked the healthcare team to enter the visit ID/QR code. This was accomplished by 3 different options: 1. typing in the visit ID number; 2. scanning the QR code with a barcode scanner; 3. if the patient had not received their visit ID/QR code, the web module would also allow the healthcare team to search by patient name and date of birth. Once entered, the web module would display required patient demographics for verification, contact information, insurance information for optional verification, and presence of previous consent for treatment on file. There were fields for entering workflow-specific comments (ex. Vehicle type) as well. For the final step of the workflow, the healthcare team would click the “Print” button. This would cause automatic printing of specimen order labels, printing of lab order requisition, print the typed patient comments directly onto the lab requisition for vehicle identification, complete and document a verbal consent to treat if needed, create the new central lab encounter, auto-activate the order for lab processing, and auto-generate a patient/parental/minor work and school excuse letter accessible within our EMR and patient portal. By distilling approximately 10 different steps into a single clickable button, efficiency of curbside collections was transformed, and the patient experience of collections improved dramatically.

Patient Self-Directed Screening Questionnaire

Despite these significant advances in ordering processes, on-site collection workflows, and lab processes, there remained an appetite for improving the patient experience of screening even further. We initially had leveraged a standalone chat-bot style screening tool on our patient-facing website, which would indicate to the patient whether or not they met testing criteria. However, after completing the screening, the patient would still need to interact with a member of a healthcare team (COVID-19 phone hotline, Video Visit, PCP, on-site curbside team) in order to generate an order. This step of verbally interacting with a member of the healthcare team to generate the order was unsurprisingly a large rate-limiting step. COVID-19 hotline wait times were often long and onerous. The development of an asynchronous technology-based patient-facing screening algorithm which would then map into our existing processes was highly desirable.

After several months of evaluating different options, our team released our build of an innovative patient-facing technology-based solution. A web survey tool was used to develop a questionnaire to gather patient inputted testing criteria, demographic, insurance data, offer self-scheduling options, and then automatically generate a true COVID-19 testing order within our EMR. Patients seeking COVID-19 testing were now able to go directly to our online patient self-directed screening questionnaire. If they met criteria for testing, the patient would enter their
demographic information, their insurance information, reserve an arrival window, and a true order would be generated within our EMR. The patient would then receive their visit ID/QR code via text and email within 60 seconds of completing the questionnaire.

Several key factors were leveraged for the creation of this workflow. Patient matching remains a challenge for validated and unvalidated patient registration experiences. Our workflow proposed to match patients to pre-existing records in the database using previously established algorithms. If a patient could not be matched, then a new patient record would be created.

Direct mapping of patient-entered insurance information into our Revenue Cycle software was a completely new workflow in our healthcare system. There was risk of patients choosing the wrong health insurance plan, intentionally or unintentionally. In our questionnaire, we queried the patient on several points of insurance information (unchanged, new insurance, new patient to Intermountain, no insurance) If a patient self-reported they had no health insurance, we also desired to map that to the appropriate Health Resources and Services Administration (HRSA) account for state billing. To mitigate the risk of inaccurate information mapped directly to our EMR, we created a new pandemic insurance profile in Revenue Cycle. This allowed our rules to copy pre-existing insurance information to the pandemic profile, enter new insurance to the pandemic profile, or map the HRSA account to the pandemic profile for the uninsured patient. By specifically calling out a pandemic profile, which was only populated by the patients’ use of our online screening tool, when the patient presented later for a normal non-testing related visit, the registration staff would know to re-verify insurance information.

The self-scheduling of arrival windows allowed patients to view all sites available with open time slots across the Intermountain and Idaho geography. Once the arrival window was reserved, we included an option for cancellation/rescheduling in the confirmation email, which allowed patients to be nimble in their needs while also not artificially limiting site capacity.

Reporting

As previously discussed, reporting results to patients as quickly as possible, and also reporting to public health entities was essential. Reporting to public health entities was facilitated via Electronic Lab Reporting. For delivery of results directly to patients, a people-driven process of results callbacks was the initial step. Although Intermountain Healthcare had a pre-existing robust patient portal for facilitating multiple aspects of the patient’s healthcare experience, we found that a large portion of our testing population were new patients to our healthcare system. Registering for patient portal access to view their test results initially required several steps which were challenging to a de novo patient in a contactless environment, therefore we implemented a results callback team which would call each patient with their test results. Predictably, this was an unsustainable process as the testing volumes exponentially increased. Operations decided to focus calling to only positive test results, however, even that volume became tangibly difficult to sustain with the surge in case counts. We pivoted to the concept of light identity authentication for viewing test results outside of our patient portal. While consulting closely with legal and compliance regarding patient privacy and confidentiality, we proposed a workflow of sending a unique, personalized link via text message, which when accessed by the patient, with verification of DOB, would then show the patient their result with appropriate education and guidance. Implementation of the viewing of test results after light-authentication went live in December 2020.

Results

Provider Ordering Workflow

With our novel ordering, prioritization, and lab configuration website, we have sustained >200 changes to prioritization and test codes without ever changing front-facing provider ordering workflows for over 6200 providers.

Shift from COVID-19 hotline phone screening online patient-directed screening and ordering

The capacity of our online patient-directed questionnaire to handle high volumes of screening was stable and reproducible. Maximum velocity of orders received so far has been 9 patients/minute. With the shift in order generation from requiring a patient phone call discussion with the healthcare team to our online patient-driven questionnaire, the COVID-19 hotline wait times were markedly improved. At the peak of testing demand in Fall 2020, COVID-19 hotline wait times often exceeded 1-2 hours. On 12/18/20, one day immediately prior to launching our online patient-driven questionnaire, 62% of phone calls were answered in less than 60 seconds, with a 13% abandonment rate. On 01/07/21, two weeks after deployment of our online patient-driven questionnaire, 91.2% of
Phone calls were answered in less than 20 seconds, with a 0% abandonment rate. Relative volume of inbound calls concomitantly decreased without changes in overall testing volume.

**COVID Hotline – Trends by Week (Inbound calls)**

Info, Registration, Results, Spanish Lines

![Graph of COVID Hotline Trends](image)

**Figure 2. Volumes of inbound phone calls to COVID-19 hotline from November 2020 to February 2021**

Continuous Improvement Curbside Collection time studies

After implementation of our online patient-driven screening/registration/insurance/scheduling on December 19, 2021, our on-site curbside collection patient cycle times decreased by 59%. Pre-implementation, the average patient time on site was 39.83 minutes. Post-implementation, the average patient time on site was 16.35 minutes. Our operational goal was < 20 minutes. Direct patient feedback often remarked on the improved positive experience, noting that the streamlined testing system was “fast, easy to access, with timely results.”

![Graph of Curbside Collection Times](image)

**Figure 3. On-site patient cycle times from December 2020 to January 2021**
Lab Experience time studies

In March 2020, Intermountain Central Laboratory began COVID-19 testing with the ability to process and result approximately 300 tests/day. After successful implementation of the above described measures and optimization of workflows and capacity, currently as of March 2021, Central Lab has the capacity to process and result approximately 10,000 tests/day. In summer 2020, with volumes of approximately 3000/day, often outpatient PCR turnaround times would take 4-5 days, whereas currently >97% of outpatient PCR tests are completed in less than 24 hours. As of March 10, 2021, Intermountain Central Laboratory had resulted > 1,100,000 COVID-19 tests.

Cost per test

Clinical Informatics teamed with all other partners in operations to reimagine how we could make testing more affordable, scalable, and accessible at all points of the patient experience. Digitization and automation were key contributors to decreasing the cost per test by 30.8% resulting in an estimated monthly cost avoidance of more than 2 million dollars.

Discussion

In summary, our results demonstrate how an agile, motivated, innovated Informatics team, tightly linked to operational workflow can transform healthcare. Critical to our success was the need for creativity, decisive choices, and a team mandate to never stay anchored in hubris, but to rather continually assess the problem to be addressed and create new, innovative solutions. While our team rapidly pieced together several workflows to create a highly efficient functional system, this complexity offered multiple points of failure and required constant monitoring.

Our key core functions included consumer facing data collection, stable healthcare workflows and unified screening approach for >6200 providers and screening personnel, automated flexing of ordering codes and prioritization, easily transmissible data for public health reporting, providing community testing without jeopardizing turnaround times for critically ill patients. Our testing processes truly transformed with the digitization of automated online patient-directed screening, order placement, Visit ID/QR code for collections check in, and text-message based delivery of test results.

Conclusion

There were several key points we learned: 1) Keep ordering process simple with a stable user interface; 2) Create discrete data elements of symptoms/exposure/risk; 3) Evaluate lab capacity and create automated prioritization scheme; 4) Partner with laboratory services to determine workflow, and push prioritization as a visible part of normal workflow (ex. Print priority on specimen label); 5) Adjust prioritization PRN; 6) Collaborate with operational
partners on a daily basis; 7) Use technology to improve the patient experience and create gains in testing capacity; 8) Maintain a workflow for non-tech savvy patients to mitigate possible service inequities.

These eight key points illuminate the components necessary to implement an agile, EMR agnostic, strategically delivering high-efficiency solutions for respiratory virus testing in a pandemic.

Acknowledgements

Tyler Haberle, Margit Lister, Eric Glissmeyer, RJ Bunnell, James Hellewell, Doug Nelson, Adam Kraft, Farukh Usmani, Daniel Ricks, Alex DeCleene, Trent Tuckett, Shelly Bowen, Lisa Gleed Thornton, Gary Bishop, Mel Hansen, Bill Peters, Gary Melville, Darren Mann, Pallavi Ranade, Ben Chisum, Ryan Stevenson, Alex DeCleene, Megan Smith, Gisele Borsato, Kassandra Wilson, Chris Benitez, Megan Curtis, Tony Wallin, Elise Graham, Kade Haviland, Daniel Simmonds, Susan Brown, Diane Rindlisbacher, Dave Newman, Bert Lopansri, Sarah Ilstrup, Todd Vento, Brandon Webb, Kerry Palakanis, Stirling Bennett, Karen Brownell

References

Towards more patient friendly clinical notes through language models and ontologies

Francesco Moramarco, Damir Juric, Aleksandar Savkov, Jack Flann, Maria Lehl, Kristian Boda, Tessa Grafen, Vitalii Zhelezniak, Sunir Gohil, Alex Papadopoulos Korfiatis, Nils Hammerla
Babylon Health, London, UK

Abstract

Clinical notes are an efficient way to record patient information but are notoriously hard to decipher for non-experts. Automatically simplifying medical text can empower patients with valuable information about their health, while saving clinicians time. We present a novel approach to automated simplification of medical text based on word frequencies and language modelling, grounded on medical ontologies enriched with layman terms. We release a new dataset of pairs of publicly available medical sentences and a version of them simplified by clinicians. Also, we define a novel text simplification metric and evaluation framework, which we use to conduct a large-scale human evaluation of our method against the state of the art. Our method based on a language model trained on medical forum data generates simpler sentences while preserving both grammar and the original meaning, surpassing the current state of the art.

Introduction

Making medical information available for patients is becoming an important aspect of modern healthcare, but the frequent use of medical terminology makes it less accessible for patients/consumers. There is a trade-off between promoting more “patient-friendly” medical notes\(^1\) and the efficiency of clinicians who often prefer writing in shorthand. This is an opportunity for automation, as Natural Language Processing (NLP) and Natural Language Generation (NLG) techniques have the potential to simplify medical text and thereby increase the accessibility to patients while maintaining efficiency.

Text simplification in the general domain has improved greatly with the introduction of new deep-learning methods borrowed from the field of Machine Translation\(^2\). However, the challenges in medical text simplification are particularly focused around explaining the abundant terminology, much of which is in Greek or Latin\(^3\). This is why most efforts in the field are concentrated around the use of a mapping table from complex to simple terms\(^4, 5\). While the task of language simplification is not new, there are very few datasets specifically built for it\(^6\). In the case of medical text simplification, the community has not yet been able to use a common benchmark due to data access constraints\(^7\). Perhaps, the only resource that comes close is a medically themed subset of Simple Wikipedia\(^4, 7\). In the context of clinical notes, medical accuracy and safety are of utmost importance, which makes consistent evaluation a strong requirement for sustainable improvements in the field.

We present a medical text simplification benchmark dataset of 1250 parallel complex-simple sentence pairs based on publicly available medical sample reports. Furthermore, we propose a novel approach to lexical simplification for the medical domain, which uses a comprehensive ontology of medical terms and their alternatives, and a novel scoring function that combines language model (LM) probabilities and word frequencies into one unified measure. We conduct a human evaluation to validate our method and find that unbounded, left-to-right LMs trained on medical forum data achieve the best results on our benchmark dataset. Finally, we make the source code for our method, and all materials necessary to repeat the human evaluation, available on GitHub\(^8\). While evaluated in the medical domain, this approach can be abstracted into other domains by utilising an appropriate alternative ontology and suitable language model training data.

Our contributions are the following: a dataset of simplified medical sentences, a new approach for text simplification, an evaluation framework for text simplification, and a model that generates simpler, grammatically correct sentences with their original meaning preserved.

\(^1\)https://github.com/babylonhealth/laymaker
Related work

**General text simplification.** Initial efforts on automatic text simplification use Phrase-based Machine Translation (PB-MT) methods driven by the availability of two resources: the open-source framework Moses and the Simple English Wikipedia dataset. These early PB-MT systems perform well, but remain too careful in suggesting simplifications. Later work provides extensions that address some of these issues — deletion and Levenshtein distance based ranking. Stajner et al. (2015) provide an insight into how much of an effect the size and the quality of the training data has on the performance of the MT systems.

Machine translation algorithms trained on parallel monolingual corpora, such as the Newsella parallel corpus, have shown great promise in recent years, combining, ideally, lexical and syntactic simplification. Nisioi et al. (2017) use the OpenNMT package to simultaneously perform lexical simplification and content reduction. Sulem et al. (2018b) show that performing sentence splitting based on automatic semantic parsing in conjunction with neural text simplification (NTS) improves both lexical and structural simplification.

**Medical text simplification.** A complex vocabulary is typically the main hindrance to understanding medical text, and is therefore the main target for simplification. Fortunately, there are numerous medical ontologies containing multiple ways of expressing the same medical term, often including an informal, layman alternative. Using these ontologies to replace complicated words with more common ones is a recurring theme in medical text simplification. Abrahamsson et al. (2014) show a preliminary study on a method that replaces specialised words derived from Latin and Greek with compounds from everyday Swedish words, and achieve encouraging results on readability. Shardlow et al. (2019) use existing neural text simplification software augmented with a mapping between complex medical terminology and simpler vocabulary taken from the alternative text labels of SNOMED-CT. Their simplification method has an increased understanding among human evaluators based on a crowd-sourced evaluation process. Van den Bercken et al. (2019) use a neural machine translation approach that is aided by a terminology-mapping table that decreases the medical vocabulary in the (complex) source text.

Despite these efforts, the field still lacks a benchmark dataset based on real medical data as well as accessible open source medical baselines; the exception being the small, medically themed subset of Simple Wikipedia provided by Van den Bercken et al. (2019). The main drawback of this corpus is that it tends to simplify sentences by omitting some of the information, which is not a viable method in the context of clinical notes. Medical data is highly sensitive and even its use for research purposes is strictly regulated and often difficult. Therefore a new medical data resource is bound to have a great impact and move the field forward, as it has happened in the past.

**Dataset**

The MTSamples dataset comprises around 5,000 sample medical transcription reports from a wide variety of specialties uploaded to a community platform website. However, publicly available annotations are limited to only include high-level metadata, e.g. the medical speciality of a report.

We create a parallel corpus of clinician-simplified medical sentences on the basis of the raw MTSamples dataset. We pre-process the entire original dataset by tokenising all sentences and expanding abbreviations based on a custom list of common medical ontologies compiled by clinicians. We then review and exclude sentences that have too little context (i.e. are confusing or ambiguous to a clinician) or grammatically incorrect. Finally, three clinicians (native British English speakers) create a new version of each sentence using layman terms, ensuring consistency of both structure and medical context and accuracy. Only one simple sentence is generated for each original sentence for which simplification is possible. The resulting dataset contains 1,250 sentence pairs, of which 597 (47.76%) have been simplified. The remainder have been left unchanged because they could not be further simplified. The average number of tokens in the original sentences is 66.96, and in the simplified sentences 68.60.

We divide the data into a 250 sentence development set and a 1,000 sentence test set.

---

1. Abrahamsson et al. (2014)
2. https://newsela.com/data
3. https://mtsamples.com
Medical ontologies

Recognising concepts and subsequently linking with a medical ontology are common medical NLP tasks necessary for higher level analysis of medical data\textsuperscript{24}. Many semantic tagging systems use the labels (text representations) defined as part of the ontologies to recognise possible instances of the entities in the text. Typically, every concept has a primary official label as well as at least a few alternative labels. Ideally these labels should be interchangeable; thus, they can be used to replace more complicated labels with layman alternatives.

In order to maintain good coverage of both medical terms and layman terminology, we select three state-of-the-art medical ontologies for creation of our phrase table. SNOMED-CT is one of the most comprehensive medical terminologies in the world, and is also available in different languages. As of the January 2019 release, it comprises 349,548 medical concepts, covering virtually all medical terminology used by clinicians. We also include the Consumer Health Vocabulary (CHV), the purpose of which is lexical simplification\textsuperscript{25}, and the Human Phenotype Ontology (HPO), which is a standardized vocabulary of phenotypic abnormalities encountered in human disease, and also contains a layer of plain language synonyms\textsuperscript{26}.

We create a vocabulary of medical terms (named entities) based on the labels of concepts from these ontologies — approximately 460,000 labels from 160,000 concepts. For example, the concept label “Otalgia” has alternative labels “Pain in ear” (Snomed), “Earache” (CHV), and “Ear pain” (HPO). To produce it, we align the ontologies using the union-find algorithm\textsuperscript{27} and discard duplicate labels, as well as those without alternatives, as they cannot contribute to the simplification process.

Lexical simplification

Lexical text simplification looks to identify difficult words and phrases and replace them with alternatives based on some measure of simplicity. Word frequency over a large amount of text is often chosen as this measure and has been used to both identify and replace candidates\textsuperscript{4}. The probability score of a sentence based on some language model has also been used to rank candidates\textsuperscript{28}. Additionally, in the medical domain, terminology words are often assumed to be the main target of lexical simplification\textsuperscript{4, 5}. We propose a new approach to medical lexical text simplification, which uses a vocabulary based on a medical ontology (see Section ) to identify candidates. It then ranks each alternative using a linear combination of word frequency and the sentence score produced by a language model. After completing the replacement and ranking steps for each medical term (of one or more words) in an input sentence, the process is repeated until no further changes are suggested. Figure 1 shows a high-level view of the algorithm.

Candidate ranking

The main task of lexical text simplification is to make the overall sentence simpler, so a ranking function should aim to provide the simplest replacement for each entity. However, this introduces a second challenge – maintaining correct grammar after the replacement. A good ranking function should therefore optimise for both the simplicity and grammaticality of the result.

Word frequency is a strong indicator for simplicity\textsuperscript{29} as it directly measures how common a given word is. However, there are different approaches to how it is utilised for multi-word expressions. Common approaches include taking the average\textsuperscript{5}, the median, or the minimum word frequency. We choose the minimum, under the assumption that the least frequent word in the sequence drives the overall understanding of the sequence. For example, consider the candidates

\begin{figure}
\centering
\includegraphics[width=\textwidth]{simplification_algorithm.png}
\caption{A flow diagram of the simplification algorithm.}
\end{figure}
otalgia of ear, and earache. An average or median frequency would score option 1 as simpler because of the very common word of, whereas the minimum word frequency would score option 2 as more frequent. To calculate the word frequencies (WF) for a given set of candidate labels, we use the wordfreq python package which provides word frequency distributions calculated over a large general purpose corpus.

\[
P(w) = \frac{C(w)}{|W|}
\]

where \(C(w)\) is the number of times the word occurs in the corpus and \(|W|\) is the number of words in the corpus. Given a sequence, such as heart attack, composed of words \(w_1 \ldots w_k\) we calculate its word frequency score, \(WF\), as

\[
WF(w_1 \ldots w_k) = \min_{i=1}^{k} \ln (P(w_i) + \epsilon)
\]

As we seek to combine this probability with language model scores, it makes sense to convert it to a logarithmic scale to avoid computational underflow. For the same reason, we introduce Laplace smoothing through the addition of the constant \(\epsilon (10^{-10})\).

Language models have made impressive strides in recent years, showing that they are capable of generating complex syntactic constructions while maintaining good grammar and coreference. We argue that the latter quality makes them a good predictor of grammatical correctness. Given that lexical simplification relies on the replacement of a recognised span from the sentence with a simpler one from a vocabulary, language models can be used to determine a score for how well a new term conforms to the grammar of the sentence.

To calculate this score we train a language model on a dataset of 160,000 original, top-level posts (1.8M sentences), scraped from the Reddit’s AskDocs forum. This dataset contains sentences which are largely medical and therefore will have the necessary vocabulary, while its language style is predominantly layman since the top post in a thread is usually written by a non-expert looking for medically-related information.

Given a sequence of words \(w_1 \ldots w_n\) and a language model, we can estimate the likelihood of the sequence as the log-probability of each word occurring given all preceding words in the sentence:

\[
\ln \hat{P}(w_1 \ldots w_n) = \frac{1}{n} \sum_{i=1}^{n} \ln P(w_i|w_{i-1}, \ldots, w_1, \langle s \rangle)
\]

where \(\langle s \rangle\) is the start symbol and \(n\) the number of tokens in the sequence. We normalise by the number of tokens \(n\) to account for replacement terms of different length, e.g. dyspnoea and shortness of breath. The language model gives a signal for how appropriate and grammatically correct the replacement term in the given sentence is. Table 1 shows both the language model \(LM\) and frequency \(WF\) scores for the term replacements of myocardial infarctions. In our example, the \(LM\) scores heart attacks (notice the plural) above heart attack given the context Patient had multiple.

Given the frequency score \((WF)\) of a replacement term \((T_i)\) and the language model score \((LM)\) for its corresponding replacement sentence \((S'_i)\), we define the final score as a linear combination of the two:

\[
\text{Score}(T_i) = \alpha LM(S'_i) + (1 - \alpha) WF(T_i)
\]

We then select the term with the highest score. The parameter \(\alpha \in [0, 1]\) acts as a regulariser and can be fine-tuned on a separate dataset. When \(\alpha = 0\), the score is entirely driven by \(WF\). When \(\alpha = 1\), the score is entirely driven by \(LM\). We select suitable \(\alpha\) values on the development set.

Simplification algorithm

A comprehensive vocabulary often results in overlapping candidate spans. For example, in the sentence Patient has lower abdominal pain, the following 5 spans match an entity: lower, abdominal pain, abdominal, pain, and lower.

\[\text{http://pypi.org/project/wordfreq/2.2.1}\]

\[\text{https://www.reddit.com/r/AskDocs/}\]
Table 1: LM and Frequency scores for alternative labels of myocardial infarctions.

<table>
<thead>
<tr>
<th>Candidate</th>
<th>LM Score</th>
<th>Freq. Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient had multiple...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>myocardial infarctions</td>
<td>-5.45</td>
<td>-14.32</td>
</tr>
<tr>
<td>heart attack</td>
<td>-4.38</td>
<td>-9.05</td>
</tr>
<tr>
<td>heart attacks</td>
<td>-3.91</td>
<td>-9.05</td>
</tr>
<tr>
<td>mies</td>
<td>-6.09</td>
<td>-14.34</td>
</tr>
<tr>
<td>myocardial necrosis</td>
<td>-6.13</td>
<td>-14.23</td>
</tr>
</tbody>
</table>

Table 2: An example of the language model (LM) convergence.

<table>
<thead>
<tr>
<th>Iteration</th>
<th>Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>hyperlipidemia with elevated triglycerides.</td>
</tr>
<tr>
<td>Iteration 1</td>
<td>elevated lipids in blood in addition to high triglycerides.</td>
</tr>
<tr>
<td>Iteration 2</td>
<td>excessive fat in the blood with high triglycerides.</td>
</tr>
</tbody>
</table>

abdominal pain. In the case where two or more spans overlap or one is subsumed by the other, the algorithm takes a greedy left-to-right processing approach. It ranks the spans in order from left to right, prioritising longer spans and ignoring all spans that have any overlap with an already processed span. Additionally, it is fairly common for a sentence to contain more than one non-overlapping medical terms. For example, consider the artificial sentence: *Patient has a history of myocardial infarction, tinnitus, otalgia, dyspnoea and respiratory tract infection.*, which has multiple, non-overlapping spans suitable for replacement. When constructing candidate sentences to score, replacing only one complex term while leaving the rest of the sentence unchanged yields a sub-optimal score. The optimal approach would be to perform an exhaustive search of all possible combinations within the sentence. Given $n$ terms, and $r$ replacements per term on average, exhaustive search would require $r^n$ combinations, i.e. exponential in the number of terms in the sentence. Rather than introduce this computational cost, we instead consider each term independently of the others. After simplifying all of them, we repeat the extract-and-replace process (see Figure 1) until no further change occurs, i.e. until convergence (see Table 2). This reduces the time complexity to $n \cdot t$, where $t$ is the number of iterations to reach convergence. We cap the number of iterations $t$ at 5, as our experiments show only 1 out of 1000 sentences to ever reach this many iterations. In practice, we find that most sentences converge after one iteration, with a median of 1 iterations and an average of 1.19.

Experimental setup

As described in Section , our method requires a language model to score alternative terms. To assess the best model for this purpose we train three different language models. Next, we fine-tune $\alpha$ for each of them and proceed to measure their respective success against the human-generated reference. The language models we select are:

- **ngram** — a trigram language model built with KenLM\(^6\) and trained on Reddit AskDocs;
- **GPT-1** — a neural language model\(^33\) trained on Reddit AskDocs;
- **GPT-2** — a neural language model\(^32\) pretrained only on generic English text. We don’t fine-tune this model to evaluate whether general-purpose language models are better at choosing layman alternatives.

In order to evaluate our approach, we compare it against three methods from the literature:

- **NTS** — Nisioi et al. (2017)\(^15\) train an encoder-decoder on Simple Wikipedia, which contains a proportion of medical sentences;
- **ClinicalNTS** — Shardlow et al. (2019)\(^5\) augment the system by Nisioi et al. (2017)\(^15\) with a medical phrase table, which is the current state of the art for clinical text simplification;
- **PhraseTable** — a simple term replacement system based on the phrase table from Shardlow et al. (2019)\(^5\), which we consider our baseline.

The $\alpha$ parameter introduced in Equation 1 regulates the ratio of the language model and the word frequency score used for scoring a replacement term. A held-out development set of 250 sentences is used for tuning the $\alpha$ parameter for each of our models. For this purpose we use the automatic metric SARI\(^28\), as it intrinsically measures simplicity by comparing the model output against both the human reference and the input sentence. We perform grid search on the $\alpha$ space (0 to 1) for each model (see Figure 2) and select the top $\alpha$ to be used in the final evaluation.

\(^6\)https://github.com/kpu/kenlm
Traditional evaluation metrics

There are three general evaluation approaches for simplification that have been tried in the past:

- **BLEU score**\(^{34}\) is one of the standard metrics of success in machine translation and has been used in some cases for simplification\(^{35}\) as it correlates with human judgements of meaning preservation.

- **SARI** is a lexical simplicity metric that measures the appropriateness of words that are added, deleted, and kept by a simplification model\(^{4,15}\).

- Human evaluation, either through dedicated annotators or crowd-sourcing, indicating whether the generated sentences are considered simpler by the end users.

Both SARI and BLEU are intended to have multiple references for each sentence to account for syntactic differences in the simplified text. As we only have one simplified reference for each original sentence, these metrics are likely to be somewhat biased to a particular way of expression. Therefore, conducting a human annotation process can bring additional reassurance to the evaluation process.

Human annotation

We design a human evaluation process in the form of a crowd-sourced annotation task on Amazon Mechanical Turk (MTurk)\(^{36}\). The goal of the task is to determine whether a simplified sentence is better than the original. Celikyilmaz et al. (2020)\(^{37}\) identify the two most common ways to conduct human evaluation on generated text: (i) ask the annotators to score each simplified sentence independently with a Likert scale\(^{1}\), (ii) ask the annotators to compare sentences simplified by different models. We experiment with both methods and decide to opt for the latter, which produces more consistent results, as also shown by Amidei et al. (2019)\(^{38}\). For this purpose, we create sentence pairs from each original sentence (marked as A) and either a sentence simplified by the model or the gold simplification provided in the dataset (marked as B). We use the following four categories:

1. Sentence A is easier to understand.
2. Sentence B is easier to understand.
3. I understand them both the same amount.
4. I do not understand either of these sentences.

Often, the simplified sentence generated by the models is identical to the original sentence. To save annotation resources we annotate such pairs only once and extrapolate the annotation to all models. MTurk provides little control over the reading age and language capabilities of the annotators, so we have to account for some variability in the annotation. Therefore, all sentence pairs are annotated 7 times by different annotators. In total, the annotations comprise 20,965 sentence pairs derived from 2,995 unique ones. Finally, we use the option of selecting only “master” annotators\(^{7}\) for the task, as it is difficult to judge the quality of the work of particular annotators. We choose turkers

---

\(^{7}\)Master annotators are annotators whose work has not been rejected by task requesters for some period of time.
Table 3: Human judgement counts for sentence pairs from the test set for all models and the reference human sentences. S: the generated was simpler; F: the original was simpler; E: both of equal complexity; N: cannot understand either; U: was not changed by the model/human reference; SG: simplification gain as defined in Equation 2. Bold indicates best model. Scores in SG are significant ($p < 0.05$)

<table>
<thead>
<tr>
<th></th>
<th>S</th>
<th>F</th>
<th>E</th>
<th>N</th>
<th>U</th>
<th>SG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>1730</td>
<td>273</td>
<td>904</td>
<td>40</td>
<td>4053</td>
<td>0.21</td>
</tr>
<tr>
<td>n-gram</td>
<td>1452</td>
<td>1004</td>
<td>1732</td>
<td>110</td>
<td>2702</td>
<td>0.06</td>
</tr>
<tr>
<td>GPT-1</td>
<td>1404</td>
<td><strong>747</strong></td>
<td>1736</td>
<td>117</td>
<td>2996</td>
<td><strong>0.09</strong></td>
</tr>
<tr>
<td>GPT-2</td>
<td>1372</td>
<td>1077</td>
<td>1661</td>
<td>118</td>
<td>2772</td>
<td>0.04</td>
</tr>
<tr>
<td>NTS</td>
<td>587</td>
<td>855</td>
<td>1022</td>
<td>98</td>
<td>4438</td>
<td>-0.04</td>
</tr>
<tr>
<td>ClinicalNTS</td>
<td>1483</td>
<td>1597</td>
<td>404</td>
<td><strong>93</strong></td>
<td>3423</td>
<td>-0.02</td>
</tr>
<tr>
<td>PhraseTable</td>
<td><strong>2425</strong></td>
<td>2759</td>
<td><strong>269</strong></td>
<td>98</td>
<td>1449</td>
<td>-0.05</td>
</tr>
</tbody>
</table>

without medical experience, as opposed to medical professionals, because they are a good representation of the end users of such system. We assume that the human reference should both succeed more often and fail less often than any of the models. We measure the quality of the models with a Simplification Gain $SG$ that we define as the difference between successes $S$ (option 2.) and the failures $F$ (option 1.), normalised by the total number of pairs, $T$:

$$SG = \frac{S - F}{T}$$ (2)

Results

We count all judgements of the same category for each model and the human reference, and present the results in Table 3. Additionally, based on these counts we calculate the simplification gain $SG$ as described in Equation 2. We can make the following conclusions based on this data:

- the human reference is very rarely more complex than the original, which makes a considerable difference in its Simplification Gain, as opposed to most of the models, which seem to be prone to this kind of error (see columns F and SG);
- based on the Simplification Gain in SG, the GPT-1 model yields the best performance. We believe this is due to: (i) having access to the entire context (as opposed to n-gram), which makes it cautious about simplification, and (ii) being more focused on medical terminology due to its training set (as opposed to GPT-2);
- the methods we compare against have a negative Simplification Gain, meaning the number of failures exceeds the number of successes. General-purpose NTS is less eager in its simplification (column U in Table 3), which could be explained by the divergence between its training set (Simple Wikipedia) and our test set (Clinical Notes). Both ClinicalNTS and PhraseTable overcome this by applying a medical phrase table (see Section 6 for more details), which triggers more medical replacements. ClinicalNTS has higher Simplification Gain overall compared to general purpose NTS, which is to be expected, but still fails more often than succeed;
- A possible explanation for the high number of successes of NTS and ClinicalNTS lies in their aggressive removal of phrases, which makes them easier to understand, but at a considerable loss of information. Both systems use a model trained on Simple Wikipedia, which very often simplifies sentences by removing words or phrases. For example, the original sentence: “It has normal uric acid, sedimentation rate of 2, rheumatoid factor of 6, and negative antinuclear antibody and C-reactive protein that is 7.” is simplified into “It has normal uric acid.”

We also report the scores for the most commonly used automatic metrics in the field, BLEU and SARI, though we stress that these scores are unreliable due to (i) their limitations as shown by Sulem et al. (2018) — they only use surface level syntactic features, and (ii) they perform better with multiple references and we only have one. The NTS baseline is still performing poorly in most metrics except for BLEU, which is likely due to its conservative approach resulting in a large number of unchanged sentences that likely overlap with the reference sentences.
Table 4: Reference-based metrics. **BLEU** and **SARI** calculated using the human-generated reference sentences.

<table>
<thead>
<tr>
<th>Model</th>
<th>BLEU</th>
<th>SARI</th>
</tr>
</thead>
<tbody>
<tr>
<td>n-gram</td>
<td>66.31</td>
<td>33.40</td>
</tr>
<tr>
<td>GPT-1</td>
<td>68.19</td>
<td><strong>33.57</strong></td>
</tr>
<tr>
<td>GPT-2</td>
<td>66.45</td>
<td>33.40</td>
</tr>
<tr>
<td>NTS</td>
<td><strong>70.17</strong></td>
<td>27.67</td>
</tr>
<tr>
<td>ClinicalNTS</td>
<td>68.22</td>
<td>30.14</td>
</tr>
<tr>
<td>PhraseTable</td>
<td>53.37</td>
<td>27.70</td>
</tr>
</tbody>
</table>

Table 5: Grammaticality and meaning preservation scores over a sample of 1250 generated sentences. **G1**: no errors; **G2**: minor errors; **G3**: major errors; **M**: the meaning is preserved. Bold implies best result.

<table>
<thead>
<tr>
<th>Model</th>
<th>G1</th>
<th>G2</th>
<th>G3</th>
<th>M</th>
</tr>
</thead>
<tbody>
<tr>
<td>n-gram</td>
<td>65.2%</td>
<td>25.6%</td>
<td>9.2%</td>
<td>93.3%</td>
</tr>
<tr>
<td>GPT-1</td>
<td><strong>75.4%</strong></td>
<td>17.8%</td>
<td>6.8%</td>
<td><strong>93.4%</strong></td>
</tr>
<tr>
<td>GPT2</td>
<td>69.3%</td>
<td>21.6%</td>
<td>9.1%</td>
<td>89.6%</td>
</tr>
<tr>
<td>NTS</td>
<td>75.4%</td>
<td><strong>9.6%</strong></td>
<td>15%</td>
<td>63.1%</td>
</tr>
<tr>
<td>ClinicalNTS</td>
<td>43.4%</td>
<td>42%</td>
<td>14.4%</td>
<td>60.4%</td>
</tr>
<tr>
<td>PhraseTable</td>
<td>31.6%</td>
<td>34.8%</td>
<td>33.6%</td>
<td>60.8%</td>
</tr>
</tbody>
</table>

To test the impact of convergence, we perform an ablation study on all our models. We take all the sentences that require more than one iteration to converge (around 10% of the dataset) and perform the same human annotation described in Section . Our results show that convergence improves **SG** for all models except GPT-2 by reducing the number of miss-simplified sentences. Empirically we find that the GPT-2 tends to increase the length of the sentence at each iteration, falling into a loop typical for language models.

**Grammaticality and meaning preservation**

Asking end users to rank two sentences in order of simplicity is not enough to judge whether a generative model is performing well. A model should be penalised if the simplified sentence is grammatically incorrect or if it has altered the meaning of the original sentence. To test these two criteria, we take a random sample of 1250 simplified sentences from all models from the test set. We ask a linguist to assign one of three grammaticality categories: no errors (G1), minor errors (G2), and major errors (G3). We then ask a clinician to mark sentences from the same sample where the meaning has changed in any way.

Table 5 summarises our findings. It clearly shows the contributions of a good language model in both preserving grammar and meaning. Our method, which is informed by language models, scores highest in both criteria. NTS, which uses a language model decoder, is quite successful in preserving grammar but less successful in preserving meaning. This is likely due to its training set, which encourages the model to remove complex phrases to simplify a sentence. ClinicalNTS and PhraseTable, which rely on a hard-coded phrase table of medical substitutions, score lower both in grammaticality and meaning preservation.

**Conclusion**

In this paper, we present a novel approach to medical text simplification in an effort to empower patients with valuable information about their own health.

First, we address the lack of high quality, medically accurate, and publicly available datasets for evaluating medical text simplification by creating such a dataset with the help of medical professionals. Second, we propose an evaluation framework for assessing the quality of simplification algorithms in the medical domain, including an experimental setup for crowd-sourced human evaluation and a metric, which we call Simplification Gain, to compare the outcomes. Third, we use the knowledge stored in state-of-the-art medical ontologies to construct a comprehensive ontology of alternative medical terms, and we develop a method for simplifying medical text by extracting and replacing medical terms with layman alternatives. To rank the alternatives, we define a scoring function that takes into account both the frequency of the replacement term and how well it fits into the sentence. Our experiments, using crowd-sourcing, show that our method is capable of simplifying complex medical text while retaining both its grammatically and meaning.

We show that our method surpasses the state-of-the-art systems in medical text simplification, improving on grammaticality and meaning preservation of the simplified sentences. These aspects are particularly important in the context of medical text simplification, where factual correctness is paramount.

888
References

1. Academy of Medical Royal Colleges. Please, write to me. writing outpatient clinic letters to patients. 2018.


First-line drug resistance profiling of Mycobacterium tuberculosis: a machine learning approach

Stephanie J. Müller, PhD1, Rebone L. Meraba, MSc1, Gciniwe S. Dlamini, MSc1, Darlington S. Mapiye, PhD1
1IBM Research Africa, Johannesburg, South Africa

Abstract

The persistence and emergence of new multi-drug resistant Mycobacterium tuberculosis (M. tb) strains continues to advance the devastating tuberculosis (TB) epidemic. Robust systems are needed to accurately and rapidly perform drug-resistance profiling, and machine learning (ML) methods combined with genomic sequence data may provide novel insights into drug-resistance mechanisms. Using 372 M. tb isolates, the combined utility of ML and bioinformatics to perform drug-resistance profiling is demonstrated. SNPs, InDels, and dinucleotide frequencies are explored as input features for three ML models, namely Decision Trees, Random Forest, and the eXtreme Gradient Boosted model. Using SNPs and InDels, all three models performed equally well yielding a 99% accuracy, 97% recall, and 99% F1-score. Using dinucleotide frequencies, the XGBoost algorithm was superior with a 97% accuracy, 94% recall and 97% F1-score. This study validates the use of variants and presents dinucleotide features as another effective feature encoding method for ML-based phenotype classification.

Introduction

Despite decades of research and pharmaceutical advancements, tuberculosis (TB) continues to affect more than 20% of the global population1. Only 30 countries account for nearly 90% of the TB cases observed worldwide, with many of these being low- and middle income countries (LMIC)1. While pharmaceutical interventions such as new anti-TB drugs remain a focus area to eradicate TB, the continuous development of drug resistance by Mycobacterium tuberculosis (M. tb) strains poses the greatest challenge in attempts to curb the epidemic.

Drug-susceptibility testing using phenotypic methods such as laboratory-based culture are both laborious and time-consuming, sometimes requiring up to three months for confirmation of drug-susceptibility of the M. tb isolate. In addition, the infrastructure and technical skills needed to perform phenotypic testing is often limited in LMIC countries which also tend to be most in need of these capabilities. To circumvent these long waiting times - during which drug-resistant M. tb continues to spread within communities - several genotypic assays have been developed for rapid screening of isolates for drug-resistance mutations2,3. Although such genotypic assays improve the turnaround time, they are limited in the mutations they screen for, thereby reducing their ability to comprehensively detect drug resistance. As a result of this design, these assays are unable to either detect mutations in other genes of the isolate, discover novel mutations arising in the bacterial community, or exclude the potential resistance to non-target M. tb drugs as caused by other mechanisms4,5.

Numerous studies have applied traditional bioinformatics methods to explore the M. tb genome and identify single nucleotide polymorphisms (SNPs) or insertions and deletions (InDels) associated with drug resistance6–8. Briefly, these studies categorized genetic variants according to those that are associated with resistance to a drug and those that are not, followed by classification using these phenotypic labels. This approach is useful for drugs such as Rifampicin (i.e Rifampin) in which resistance is conferred by a few variants with large effect size. However, its performance is reduced on drugs for which resistance is conferred by a large number of variants, possibly spanning multiple genes. Advances made in machine learning and sequencing techniques in the last decade have provided a greater platform from which to investigate and leverage the “hidden” patterns within genomic data, thereby providing an alternative to phenotypic drug resistance profiling8,9. To facilitate this, several machine learning approaches have been applied to genomic data from M. tb isolates to predict drug resistance6,9,10.

In this study, the application of machine- and statistical learning models and genomic features is presented with the aim of developing a clinically applicable model to detect extreme first-line multi-drug resistance (MDR) i.e isolates simultaneously resistant to the four first-line anti-TB drugs. We hypothesize that the predictive performance for anti-TB drugs that define extreme MDR to first-line drugs can be improved using a comprehensive set of genomic variants.
coupled with tree-based machine learning algorithms capable of incorporating and encoding them.

**Materials and Methods**

**Data collection and pre-processing**

A total of 372 *M. tb* isolates sourced from Walker *et. al* were included in this study. Here, machine learning and statistical methods were used to identify features stratifying isolates from two extreme MDR phenotypes, namely first-line drug resistance and first-line drug susceptibility. A total of 186 samples resistant to all four first-line anti-TB drugs (Rifampicin (RIF), Isoniazid (INH), Ethambutol (EMB), and Pyrazinamide (PZA)), and 186 samples susceptible to the same four anti-TB drugs were used as input to the workflow.

![Generalized workflow used in this study](image)

**Figure 1:** Generalized workflow used in this study

Using a customised bioinformatics workflow (Figure 1), all isolates were subjected to quality control, in which paired-end raw sequencing reads in FASTQ format were trimmed using Cutadapt (v. 1.18) with a Phred score of 20. The trimmed sequences were aligned to the *M. tb* H37Rv reference genome (Genbank: NC000962.3) using BWA-mem (v. 0.7.17). Samtools (v. 1.9) was used to convert Sequence Alignment Map (SAM) files to Binary Alignment Map (BAM) format. Subsequently, Picard Tools (v. 2.18.23) was used for BAM file sorting, marking of duplicate reads, adding of read groups, and indexing. SNPs and InDels were called using the Genome Analysis ToolKit (GATK, v. 3.8), followed by annotation using SNPEff (v. 4.3.1). Missense SNPs and frameshift InDels were extracted using SNPSift (v. 4.3.1).

**Variant prioritization**

Variants present in 16 *M. tb* genes with known drug resistance mutations to the four first-line anti-TB drugs were prioritized for this study, namely, rpoB for RIF resistance, abpC, fabG1, inhA, katG, ndh for INH resistance, embA, embB, embC, embR, iniA, iniC, manB, rmlD for EMB resistance, and pncA and rpsA for resistance to the drug PZA. This selection was based on the reported involvement of these genes in the occurrence of drug resistance in *M. tb*. SNP and InDel mutations not appearing in at least 99% of the 16 candidate genes were removed from further analyses. The remaining mutations were used to generate a feature matrix which served as input for training of three ML classifiers.

**Dinucleotide frequencies**

Dinucleotides are comprised of the 16 unique combinations of four nucleotides AT, AA, AC, AG, TT, TA, TC, TG, GT, GA, GC, GG, CT, CA, CC, and CG. To calculate dinucleotide frequencies, reads spanning the 16 target genes were extracted using genomic coordinates retrieved using Mycobrowser. For each sample, extracted reads were synchronized using Fastq-pair followed by assembly into contigs using the default parameters of the SPAdes assembler. Contigs are continuous sequences produced when the optimal overlap of input reads is obtained. Low read overlap, and sequence contamination may result in poor construction of a single contiguous sequence, thereby rendering either no contig or multiple shorter contigs. Samples producing a single contig and samples producing multiple contigs were selected and used further in the analysis to compute dinucleotide frequencies, where in the latter group, the longest contig was selected as recommended in Douglass *et al*. Samples producing no contigs in at
least one of the genes were excluded from further analyses as the cost of reducing the sample size was considered to be less detrimental than the cost of other measures aimed at maintaining the sample size, such as data imputation.

For a random subset of samples from each phenotype, contigs in FASTA format were submitted to the BLAST\textsuperscript{21} database via the PATRIC\textsuperscript{22} website. This was to confirm that the contigs produced co-located with the genes from which the reads were extracted. Lastly, the 16 dinucleotide frequencies of the contig sequences belonging to the genes of interest were calculated as described by Dlamini \textit{et al.}\textsuperscript{23}. To account for frequency-bias due to variable gene and contig lengths, dinucleotide frequencies were normalised across the genes for each sample to produce relative frequencies.

\textit{Data visualization}

Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor (t-SNE) were used to investigate separation of the data based on the two phenotypes. Both PCA and t-SNE may be used to reduce the dimensionality of highly sparse datasets, while maintaining the variation within the information captured in the original dataset. In addition to performing dimensionality reduction for the purposes of visualisation, PCA was performed on the feature matrix of SNPs and InDels to investigate the potential confounding effect of lineage on resistance to anti-TB drugs, as described by Gan \textit{et al.}\textsuperscript{24}. This was completed by excluding all lineage specific mutations reported in Walker \textit{et al.}\textsuperscript{4}.

\textit{Classification}

Three ML algorithms, namely, Decision Trees\textsuperscript{25} (DT), Random Forest\textsuperscript{26} (RF), and the eXtreme Gradient Boosted algorithm\textsuperscript{27} (XG) were investigated for their ability to classify the dataset into two drug resistance phenotypes, using either variants (SNPs and InDels) or dinucleotides as features. Data were classified using a binary label $Y$, with 0 used to denote drug susceptible samples, and 1 used to denote drug resistant samples. For each feature workflow, the data was randomly split into 80\% training data and 20\% test data in a stratified manner to maintain phenotype distributions. Five-fold cross-validation was performed through a randomized grid search using Scikit-learn’s RandomizedSearchCV\textsuperscript{28} to optimise the parameters of each model. Predictions were made on the held-out test set using the optimised parameters for each model. Lastly, the performance of the three models was assessed using accuracy, recall, and F1-score and 95\% confidence intervals were estimated using bootstrapping\textsuperscript{29}.

Two sets of analyses were conducted using dinucleotide features: a composite analysis (in which dinucleotide frequencies for all genes were used simultaneously as features), and a per-drug analysis, in which the classification models were retrained and tested on the subset of dinucleotide features encoding the set of genes linked to resistance of each drug. For resistance to RIF, each of the three models were trained using the dinucleotide frequencies calculated on the contig spanning the \textit{rpoB} gene. Likewise, for INH resistance, the dinucleotide frequencies calculated for all samples across the \textit{ahpC}, \textit{fabg1}, \textit{inhA}, \textit{katG}, and \textit{ndh} genes were selected, and used in model training and testing. This same process was completed for resistance to EMB and PZA, respectively, using the gene sets described previously. This analysis was performed with the goal of furthering our understanding of the mechanisms driving classification within one drug as compared to the others.

\textit{Results and Discussion}

\textit{Data collection and pre-processing}

\textit{Features}

The variant-based feature set consisted of 54 features and 372 observations, containing 186 resistant and susceptible samples each. For the dinucleotide features, contig construction was attempted for each of the 16 genes of interest and for all 372 samples. Of the 372 samples, samples producing no contig for one or more of the 16 target genes were removed from the analysis and the subset of samples producing at least one contig per gene was retrieved. This process reduced the sample size from 186 samples with resistant and susceptible phenotypes each, to 153 resistant and 182 susceptible samples, respectively. A randomly chosen set of contigs were visualized in the PATRIC database, confirming that the contigs constructed for the genes did indeed span the intended genes. In addition, more than
30% of the samples produced multiple, short contigs for the embR gene, and thus, this gene was removed from all further analyses. Therefore, the final feature dataset for dinucleotide frequencies was comprised of 335 observations (samples) and 240 features (16 dinucleotide frequencies for 15 genes).

**Data visualization**

Using variants as features, PCA and t-SNE of the 16 target genes showed a lack of multiple clusters due to a poor separation of the *M. tb* isolates by lineage (Figure 2). As expected, given the absence of the 42 lineage-defining mutations that were reported by Walker *et. al*\(^4\), these findings suggest that the samples are from a single lineage. Moreover, the lack of multiple clusters with a clear discrimination between the two phenotypes confirms that lineage is not a confounder of *M. tb* drug resistance. Relative frequencies of dinucleotides as features representing the 15 genes showed no discernible pattern of resistance classification (Figure 3).

**Figure 2:** (A) PCA and (B) t-SNE of variant features for 16 genes

**Figure 3:** (A) PCA and (B) t-SNE of dinucleotide features for 15 genes

**Classification**

When using variants as features, the XG model was trained using default parameters, while DT and RF models were trained using optimised hyperparameters. When using dinucleotide frequencies as features, all three models were trained using default values for the hyperparameters as these hyperparameter configurations yielded optimal cross-validated results.

For all three ML models, classification by drug resistance phenotype showed that using variants as features resulted in
fewer incorrect predictions (Figure 4) as compared to using dinucleotides as features (Figure 5). Although marginal, the use of variants as features performed better than using dinucleotide frequencies (Table 1). When considering the use of variants as features for phenotype classification, all models performed equally well with an average accuracy of 99%, recall of 97%, and F1-score of 99%.

When using dinucleotide features, the XG algorithm outperformed the DT and RF models, with an average accuracy of 97%, recall of 94%, and F1-score of 97% (Table 1). This may be attributed to either the input features, the existence of additional resistance-associated mutations, or the co-occurrence of resistance to multiple drugs within the genes of interest. Alternatively, it may be that the feature encoding methods used in this analysis are capable of encoding additional features associated with drug resistance.

![Confusion matrices of the variant features for three models: (A) Decision Tree, (B) Random Forest, (C) XGBoost](image1)

**Figure 4:** Confusion matrices of the variant features for three models: (A) Decision Tree, (B) Random Forest, (C) XGBoost

![Confusion matrices of the dinucleotide features for three models: (A) Decision Tree, (B) Random Forest, (C) XGBoost](image2)

**Figure 5:** Confusion matrices of the dinucleotide features for three models: (A) Decision Tree, (B) Random Forest, (C) XGBoost

<table>
<thead>
<tr>
<th>Feature encoding method</th>
<th>Model</th>
<th>Accuracy</th>
<th>Recall</th>
<th>F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Decision Tree</td>
<td>0.99 (0.96-1.00)</td>
<td>0.97 (0.91-1.00)</td>
<td>0.99 (0.95-1.00)</td>
</tr>
<tr>
<td></td>
<td>Random Forest</td>
<td>0.99 (0.96-1.00)</td>
<td>0.97 (0.91-1.00)</td>
<td>0.99 (0.95-1.00)</td>
</tr>
<tr>
<td></td>
<td>XGBoost</td>
<td>0.99 (0.96-1.00)</td>
<td>0.97 (0.91-1.00)</td>
<td>0.99 (0.95-1.00)</td>
</tr>
</tbody>
</table>

**Table 1:** Classification results: average performance of the three models with their 95% confidence intervals

A subset of important features driving the drug resistance prediction was extracted for each model. For the approach
using variants as features, the Ser315Thr mutation in the katG gene and Ser450Leu mutation in the rpoB gene were consistently identified as the top two features across all three models (Table 2). These results are similar to those reported by Walker et al., where these mutations were identified to confer resistance to INH and RIF, respectively. Similarly, in a whole genome sequencing study performed to detect genomic markers of *M. tb* drug resistance, the most common mutations observed for INH and RIF resistance were katG (Ser315Thr) and rpoB (Ser450Leu), respectively\(^30\). Furthermore, when the evaluation of genotype MTBDRplus for rapid detection of drug resistant TB was performed among resistant clinical isolates, the same mutations were dominant\(^31\).

The drug composite analysis using dinucleotide features encoding for similar samples revealed the AT dinucleotide in the iniC gene to be the top feature driving the phenotype classification for all three models (Table 2). This feature was closely followed by the TA, and AT dinucleotides in the rpoB gene driving classification using the DT, and RF models, respectively. While iniC harbours mutations linked to EMB resistance, rpoB is well-known for its mutations conferring resistance to Rifampicin\(^4\). Mutations in the fabG1 gene have also been associated with resistance of *M. tb* to INH\(^32\).

When analysing the dataset per drug, phenotype classification using dinucleotides from the rpoB gene responsible for RIF resistance performed relatively well, with a mean accuracy of 88%, 94%, and 94% for the DT, RF, and XG models, respectively (data not shown). Classification of samples into their respective phenotypes using features from the five genes associated with INH resistance also performed relatively well, with a mean accuracy of 90%, 94%, and 96% for the DT, RF, and XG models, respectively. Lastly, models built using dinucleotide features from genes linked to either PZA or EMB resistance performed worse than those built for RIF or INH (results not shown). Additionally, the TA dinucleotide in the pncA gene was found to consistently be the top performing feature driving phenotype classification using dinucleotide features from the gene set responsible for PZA resistance. Using dinucleotides from genes linked to EMB resistance, the AT dinucleotide in the iniC gene was the most important feature driving phenotype classification using the DT and XG models, while the GC dinucleotide in the manB gene was the most important feature when using the RF model.

While we are unable to validate these findings using laboratory methods at this time, it may be hypothesised that classification using dinucleotide features built using genes conferring EMB resistance in isolation may be weaker, but when analysed with features for the genes associated with the other three first-line drugs, classification using EMB features may be strengthened, as seen by the dominance of the iniC gene as a top feature (Table 2).

As shown by this study, the use of two different genomic feature encoding methods may enable the identification of different drug resistant genes. The analyses conducted has shown that although the genetic determinants of resistance to RIF and INH are well defined and understood, the full catalogue of mutations encoding resistance to other first-line drugs is not yet fully understood. Therefore, these findings demonstrate and support that rapid diagnostic tests relying on detecting mutations conferring resistance to RIF and INH are highly sensitive and specific but those targeting other drugs require further optimization if they are to replace conventional drug susceptibility testing.

Some limitations to this study include the potential error in phenotypes as drug susceptibility testing for some drugs have been shown to have low reproducibility and high variance, assuming that resistance for each drug and feature space were independent. Additionally, all variants in the genes included in this analysis were considered as equally important during model development, limiting our knowledge of strain hypervirulence\(^33\).

Whilst the effects of small sample sizes on ML algorithms are well-documented, the final models performed fairly well despite the limited dataset. This may, in part, be attributed to the type of algorithms (tree-based methods) that were implemented, whose complexity could be controlled in order to minimise over-fitting. In addition to the choice of algorithm, the feature set that was used to fit the models also contributed to the models’ performance. In situations where the dataset is limited, it is necessary to utilise knowledge-based features that are known to have a direct influence on the prediction task. In this study, this manifested in two ways. Firstly, it was through the focus on the 16 genes that have been implicated in drug resistance in the four first-line drugs. Secondly, and specifically for the variant-based analysis, it was through the additional employment of hand-crafted, expert-informed variant features that were extracted from the 16 genes of interest.
Table 2: Top ten features for each model ranked in order of descending importance for each feature encoding method

<table>
<thead>
<tr>
<th>Feature encoding method</th>
<th>Decision Tree</th>
<th>Random Forest</th>
<th>XGBoost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variant calling</td>
<td>katG_Ser315Thr</td>
<td>katG_Ser315Thr</td>
<td>katG_Ser315Thr</td>
</tr>
<tr>
<td></td>
<td>rpoB_Ser450Leu</td>
<td>rpoB_Ser450Leu</td>
<td>rpoB_Ser450Leu</td>
</tr>
<tr>
<td></td>
<td>embR_Cys372Gly</td>
<td>katG_Arg463Leu</td>
<td>embB_Met306Val</td>
</tr>
<tr>
<td></td>
<td>embC_Arg738Gln</td>
<td>embB_Met306Val</td>
<td>iniA_His481Gln</td>
</tr>
<tr>
<td></td>
<td>embC_Val981Leu</td>
<td>embR_Phe376Leu</td>
<td>embB_Pro12Gln</td>
</tr>
<tr>
<td></td>
<td>embR_Phe376Leu</td>
<td>embB_Met306Ile</td>
<td>manB_Asp152Asn</td>
</tr>
<tr>
<td></td>
<td>katG_Arg463Leu</td>
<td>embR_Cys372Gly</td>
<td>embR_Phe376Leu</td>
</tr>
<tr>
<td></td>
<td>pncA_Ile133Thr</td>
<td>embC_Arg738Gln</td>
<td>katG_Arg463Leu</td>
</tr>
<tr>
<td></td>
<td>embB_Met306Val</td>
<td>pncA_Gln10Pro</td>
<td>embC_Val981Leu</td>
</tr>
<tr>
<td></td>
<td>embB_Met306Ile</td>
<td>pncA_Ile133Thr</td>
<td>embR_Cys372Gly</td>
</tr>
</tbody>
</table>

Dinucleotide features

<table>
<thead>
<tr>
<th></th>
<th>iniC_AT</th>
<th>iniC_AT</th>
<th>iniC_AT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>rpoB_TA</td>
<td>rpoB_TA</td>
<td>fabG1_AT</td>
</tr>
<tr>
<td></td>
<td>inhA_AT</td>
<td>fabG1_AT</td>
<td>fabG1_AT</td>
</tr>
<tr>
<td></td>
<td>ahpC_AA</td>
<td>iniA_CG</td>
<td>embC_TA</td>
</tr>
<tr>
<td></td>
<td>ndH_TG</td>
<td>pncA_TA</td>
<td>rpoB_TA</td>
</tr>
<tr>
<td></td>
<td>iniA_CA</td>
<td>rpoB_TA</td>
<td>manB_GC</td>
</tr>
<tr>
<td></td>
<td>katG_GT</td>
<td>embC_AT</td>
<td>pncA_AA</td>
</tr>
<tr>
<td></td>
<td>katG_AC</td>
<td>embC_GC</td>
<td>fabG1_TC</td>
</tr>
<tr>
<td></td>
<td>iniC_AA</td>
<td>ahpC_TA</td>
<td>inhA_AT</td>
</tr>
<tr>
<td></td>
<td>ndH_AA</td>
<td>manB_GC</td>
<td>ahpC_TA</td>
</tr>
</tbody>
</table>

Conclusion

In this study, three ML models were evaluated for their ability to classify the phenotype of *M. tb* samples as being resistant or susceptible to the four first-line anti-TB drugs. These models used two different feature encoding methods. The use of variants (SNPs and InDels) as features performed slightly better than the models constructed using dinucleotide frequencies as features. Furthermore, a review of the model features that drove the classification results for the variant-based analysis revealed a set of known important features, thereby confirming the utility of this feature encoding approach. In contrast, the important features driving classification using dinucleotide features provided unexpected results, with the AT dinucleotide in the iniC gene (associated with EMB resistance) being the top performing feature. This study therefore demonstrates the potential use of different genomic variation encoding methods combined with machine learning in clinical workflows to robustly predict anti-TB drug resistance. The results obtained in this study demonstrate the potential of phenotypic drug resistance profiling through encoding underlying genomic variation into features that can be used to develop machine learning models capable of discriminating *M. tb* isolates with extreme forms of resistance.
References


Recurrent Neural Network based Time-Series Modeling for Long-term Prognosis Following Acute Traumatic Brain Injury

Amin Nayebi1, Sindhu Tipirneni2, Brandon Foreman3, Jonathan Ratcliff4, Chandan K Reddy2, Vignesh Subbian1

1The University of Arizona, AZ, USA; 2Virginia Tech, VA, USA; 3University of Cincinnati, OH, USA; 4Emory University, GA, USA

Abstract

We developed a prognostic model for longer-term outcome prediction in traumatic brain injury (TBI) using an attention-based recurrent neural network (RNN). The model was trained on admission and time series data obtained from a multi-site, longitudinal, observational study of TBI patients. We included 110 clinical variables as model input and Glasgow Outcome Score Extended (GOSE) at six months after injury as the outcome variable. Designed to handle missing values in time series data, the RNN model was compared to an existing TBI prognostic model using 10-fold cross validation. The area under receiver operating characteristic curve (AUC) for the RNN model is 0.86 (95% CI 0.83-0.89) for binary outcomes, whereas the AUC of the comparison model is 0.69 (95% CI 0.67-0.71). We demonstrated that including time series data into prognostic models for TBI can boost the discriminative ability of prediction models with either binary or ordinal outcomes.

Introduction

Traumatic Brain Injury (TBI) is one of the leading causes of death and disability in the United States. There are nearly 2.8 million new TBI cases every year in the US (1). It is estimated that the general trend of worldwide TBI cases will continue to increase and be a growing significant health problem, particularly, for low- and middle-income countries (2). Accurate and early prediction of outcomes in TBI patients can help in clinical management as well as in optimizing resource allocation within the health system (3).

Even though not all TBI cases result in death, different forms of disability are common in complex TBI cases, and simple mortality predictions do not account for the long-term health consequences associated with TBI. Therefore, any practical TBI prediction model should consider outcomes other than mortality (2). The Extended Glasgow Outcome Scale (GOSE) is a functional TBI outcome measure of the severity of TBI and is often used in prognostic models (4). GOSE rates patients in eight categories, from death to upper good recovery, and has been commonly dichotomized into mortality (versus survival) or unfavorable (versus favorable). In this research, we develop a model to predict the GOSE outcome of TBI patients.

Prior studies have developed and evaluated models to predict mortality or severity of TBI patients. A systematic review shows that many prognostic studies suffer from methodological issues. For example, the sample size may not be sufficient: 75% of studies have less than 500 subjects (2,3,5). Two of the most widely used prediction studies for TBI patients are the International Mission on Prognosis and Analysis of Clinical trials in Traumatic Brain Injury (IMPACT) (6) and Corticosteroid Randomization After Significant Head injury (CRASH) (7). The covariates in these models are primarily based on the clinical, physiological, and lab data that are collected at the time of admission, which does not take into account the evolution of the primary injury, and the development of secondary brain injuries.

Regression models, such as logistic regression, to predict disease occurrence (diagnosis) or disease outcome (prognosis), are standard approaches to build a prediction model (8). However, machine learning (ML) algorithms are gaining acceptance for use in the clinical domain especially as the increasingly large and rich data sets such as Electronic Health Records (EHR) data are growingly available (9). On the other hand, recent data suggests that ML algorithms do not necessarily outperform regression models for prognosis of TBI cases, especially when the number of predictors is not high.

In our study, in addition to clinical and laboratory data collected at the time of admission to the Emergency Department (ED), we use time series data obtained from the first few days in the ICU. Recurrent neural networks (RNN) are well known to achieve strong results in many applications with time series and sequential data (10). Two common RNN structures, the Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU), can capture the long-term temporal dependencies in variable-length samples. GRUs are getting more attention since they can maintain the effect of LSTM units while they are simpler. A study showed that GRUs can outperform LSTM units both in terms of CPU time and generalization (11).

One of the limitations in working with clinical time series data is missingness. Approaches to handle missing values in time series data include, but are not limited to, deletion, mean imputation, and autoregression (12). A case study shows that when the missing rate is high, excluding incomplete data negatively impacts the performance of prediction models compared to alternative scenarios of imputation (13). In this work, imputing time series missingness and training the model are done simultaneously. GRU-D is a recently developed recurrent unit for managing the missingness that optimizes the model performance by imputing missing data while simultaneously learning the model parameters (14).
In this paper, we describe an attention-based RNN with new recurrent units known as GRU-D. Our study differentiates from others according to the following characteristics.

- The model combines the temporal features during the ICU stay (e.g., a sequence of vital signs) and non-temporal features such as age and sex.
- The model handles missing values in the time series data and learns from the missingness patterns. Training and imputation occur simultaneously. It uses GRU-D units that benefit from a decaying mechanism for imputation of missing values among time series data.
- The most important features for predicting the longer-term outcomes in acute TBI patients are identified.

Materials and methods

Source of data

This study was based on data from the prospective, multicenter Transforming Research and Clinical Knowledge in Traumatic Brain Injury (TRACK-TBI) study (15). The TRACK-TBI study collected detailed clinical data on TBI patients from 18 different academic Level I trauma centers across the US. TRACK-TBI enrolled 2996 participants across the spectrum of TBI severity.

For this analysis, we included a total of 110 clinical variables that were collected at the time of arrival to the ED, discharge from ED, and during the ICU or hospital stay. During ICU stays, measurements were recorded as frequently as every hour, providing extensive time-series data for those variables. Of the total 110 variables, 59 were static variables (demographic or one-time-recorded measurements in ED), and 51 were time-series variables that were recorded during the patients hospitalization. The GOSE was measured at six months after injury and was used as our primary outcome variable for this analysis. Table 1 shows the summary statistics for clinical variables in the dataset.

Table 1. Summary statistics of clinical variables in TRACK-TBI dataset

<table>
<thead>
<tr>
<th>Variable</th>
<th>Frequency/mean</th>
<th>Percentage</th>
<th>% Missing data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demographic</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age (mean +/- SD)</td>
<td>39.11 +/- 18.25</td>
<td>31.5%</td>
<td>0%</td>
</tr>
<tr>
<td>Sex (female)</td>
<td>881</td>
<td></td>
<td>0%</td>
</tr>
<tr>
<td>ED examination</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GCS (mean +/- SD)</td>
<td>12.95 +/- 3.93</td>
<td></td>
<td>4.9%</td>
</tr>
<tr>
<td>Pupil Reactivity</td>
<td></td>
<td></td>
<td>18.8%</td>
</tr>
<tr>
<td>Both</td>
<td>2124</td>
<td></td>
<td>93.4%</td>
</tr>
<tr>
<td>Neither</td>
<td>118</td>
<td></td>
<td>5.2%</td>
</tr>
<tr>
<td>One</td>
<td>32</td>
<td></td>
<td>1.4%</td>
</tr>
<tr>
<td>Motor Score</td>
<td></td>
<td></td>
<td>4.9%</td>
</tr>
<tr>
<td>No response</td>
<td>225</td>
<td></td>
<td>8.4%</td>
</tr>
<tr>
<td>Extension</td>
<td>25</td>
<td></td>
<td>0.9%</td>
</tr>
<tr>
<td>Abnormal</td>
<td>22</td>
<td></td>
<td>0.8%</td>
</tr>
<tr>
<td>Withdrawal</td>
<td>77</td>
<td></td>
<td>2.9%</td>
</tr>
<tr>
<td>Localize</td>
<td>140</td>
<td></td>
<td>5.2%</td>
</tr>
<tr>
<td>Obey</td>
<td>2134</td>
<td></td>
<td>80.1%</td>
</tr>
<tr>
<td>Untestable</td>
<td>40</td>
<td></td>
<td>1.5%</td>
</tr>
<tr>
<td>Diastolic Blood Pressure (mean +/- SD)</td>
<td>84.1 +/- 18.3</td>
<td></td>
<td>6.0%</td>
</tr>
<tr>
<td>Systolic Blood Pressure (mean +/- SD)</td>
<td>139.9 +/- 24.2</td>
<td></td>
<td>1.4%</td>
</tr>
<tr>
<td>Hemoglobin (mean +/- SD)</td>
<td>13.9 +/- 1.7</td>
<td></td>
<td>12.0%</td>
</tr>
<tr>
<td>Glucose (mean +/- SD)</td>
<td>134.7 +/- 53.0</td>
<td></td>
<td>13.1%</td>
</tr>
<tr>
<td>Complications and treatment</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre-hospital Hypotension (yes)</td>
<td>91</td>
<td></td>
<td>3.3%</td>
</tr>
<tr>
<td>Pre-hospital Hypoxia (yes)</td>
<td>77</td>
<td></td>
<td>2.8%</td>
</tr>
<tr>
<td>6-month Outcome</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GOSE</td>
<td>302</td>
<td></td>
<td>37.4%</td>
</tr>
<tr>
<td>1- Death</td>
<td>126</td>
<td></td>
<td>7.2%</td>
</tr>
<tr>
<td>2- Vegetative state</td>
<td>6</td>
<td></td>
<td>0.3%</td>
</tr>
</tbody>
</table>

901
3- Lower severe disability  88  5.0%
4- Upper severe disability  23  1.3%
5- Lower moderate disability  160  9.1%
6- Upper moderate disability  304  17.3%
7- Lower good recovery  464  26.5%
8- Upper good recovery  582  33.2%

Study participants
We included all adult participants with GOSE scores available and those that are admitted to the ICU. Non-adult participants and those who withdrew consent are excluded from the analysis. Out of 2996 participants, 902 met the mentioned criteria and are included in the analysis. Only the first five days of ICU data are used in this study. Figure 1 shows the exclusion criteria and the number of subjects that are left after applying each criterion.

Figure 1. Flowchart showing study participant selection

Missing data
Large amount of missing values in clinical time series data is common (16). To tackle sporadic measurements in the ICU, we discretized the observation time window into fixed-length time intervals. As shown in Figure 2, some variables might have missing values after discretizing the time window, while others have more than one value in a timestamp. We replaced the value of each variable in each timestamp with the average of recorded measurements in the corresponding interval. In order to handle the missingness, we utilized a modified version of Gated Recurrent Units which will be discussed later.

From all static variables, those with more than 20 percent missing data were excluded from the analysis. For imputing the missing static data, we used a Multivariate Imputation with Chained Equations (MICE) approach. In this approach, a series of predictions are used to impute the missing values of each variable. This is done iteratively until the imputed data does not change significantly (17).

Model Development
In this study, a deep RNN model was developed to predict TBI patients’ functional outcome at six months, post injury. As the output of the model (GOSE) is an ordinal variable, we follow the same procedure presented by Chen et al. (18) to handle the ordinal output in a neural network. We transformed a classification problem with $K$ ordinal categories to $K-1$ binary classifications. To do so, we used a special type of output encoding. The $i^{th}$ element of the encoded binary output shows whether the original output is larger than the $i^{th}$ ordinal level. In other words, if a data point belongs to the $i^{th}$ category, the first $i-1$ binary variables of the encoded output vector are 1 and the rest are 0. As an example, if GOSE value is 3, the corresponding vector is $(1,1,0,0,0,0,0)$. We use the sigmoid function as the activation function of the output node, and a squared error loss function is used.
Figure 2. Distribution of recorded measurements across different variables over time for a single patient. The plot on the left shows that different variables have different frequencies of measurement. For example, blood pressure is taken hourly, but urea lab is taken at most once a day. The plot on the right illustrates how values are aggregated over 10-hour intervals.

The model consists of two parts. In the first part, time series data for the first five days of ICU is fed to an RNN which is elaborated on later. The output vectors of RNN at each time step go through an attention layer which takes a weighted sum of its input vectors. The attention layer output is concatenated with the static values and passed through a hidden dense layer before the output dense layer. Figure 3 shows the structure of the prediction model.

Figure 3. The deep RNN structure for predicting the output. Temporal features are fed to the GRU units and static features are connected to the output through a hidden layer. An attention layer is also used on top of the sequential part of the model.

In this model, we used a modified GRU unit presented by Che et al. (14) which is called GRU-D. They showed that the missingness pattern among variables in time series can provide useful information. They modified the GRU unit in such a way that it captures the missingness information and, at the same time, imputes the missing data. In this research, we used the GRU-D units for RNN cells in our model to predict the GOSE among TBI patients.

In the clinical domain, variables tend to be close to a default number when they are unobserved for a long time (14). This would mean that missing values in time series would fade gradually to a default value (e.g., the empirical average of the variable). GRU-D enjoys a decay mechanism for input variables and hidden states to address the mentioned properties. Two types of decay variables are used in GRU-D, input decays ($\gamma_x$) and hidden state decays ($\gamma_h$). The general formulation of decay vectors is as follows:

$$\gamma_t = \exp\{-\max(0, W_\gamma \delta_t + b_\gamma)\}$$

(1)

where $\delta_t$ shows how far the last observation of each variable is from time $t$. $W_\gamma$ and $b_\gamma$ are the parameters of the model which must be learned. Corresponding to each variable, a masking vector, $m$, is defined in such a way that is 1 if the variable is observed and otherwise is 0. To impute the input values of the time series, the following formulation is used.

$$\hat{x}_t^d = m_t^d x_t^d + (1 - m_t^d)(\gamma_x^d x_t^d + (1 - \gamma_x^d)\hat{x}_t^d)$$

(2)
In Equation (2), indices $d$ and $t$ indicate $d$-th variable and $t$-th time slot, respectively. $\hat{x}_t^d$ is the imputed value of the input variable and would be used in the traditional GRU equations. $x_t^d$ is the input time series data. $\gamma_t^d$ is the input decay calculated by Equation (1). $\hat{x}_{t'}^d$ is the last observed value of $d$-th variable at time $t'$. $\bar{x}_t^d$ is the empirical average of $d$-th variable over all time steps and observations.

To fully capture the missingness information, a decaying mechanism for the hidden states is utilized. Before modifying each hidden state using traditional GRU equations, a new decayed hidden state is calculated as follows:

$$\hat{h}_{t-1} = y_{ht} \odot \hat{h}_{t-1}$$

(3)

Instead of $x_t$ and $h_{t-1}$, we use $\hat{x}_t$ and $\hat{h}_{t-1}$ in the update functions of the GRU-D.

To avoid our model overfitting on the training data, we used some tools to regularize the model. We utilized dropout after each layer of the model (both recurrent and feed-forward layers) to randomly drop some of the nodes. We also applied a L2-regularizer on the hidden layer weight parameters. L2-regularizer applies penalties on the layer weight parameters. These penalties are summed into the loss function and hence, avoids weight parameters taking large values.

For comparison purposes, we developed a regression model using static data. In this model, only the variables used in the IMPACT prediction model (6) are included, which are age, motor score, pupillary reactivity, hypoxia, hypotension, glucose, and hemoglobin. To build the model, we first converted the values of each variable to the IMPACT score, and then developed an ordinal logistic model on those scores. The IMPACT model originally was developed on patients with moderate and severe TBI (GCS ≤ 12), and a dichotomized GOS score based on favorable and unfavorable outcomes. To have a fair comparison with IMPACT, we also developed, trained, and validated our model under the same conditions. In order to dichotomize the GOSE score, an outcome is unfavorable if GOSE ≤ 4, otherwise it is considered favorable.

The output of the prediction is an ordinal variable, and generally three types of performance metrics are used to assess ordinal classifiers: accuracy, misclassification error, and rank association (19). All these three measurements are used in this study to compare the models. Accuracy (ACC) simply calculates the proportion of correct classifications. To take the misclassification into account, we also used Area Under the Curve (AUC), F1 score, and Mean Squared Error (MSE). MSE measures the degree of error between true and predicted labels. To calculate this error, we assigned a number to each class of GOSE score from one to eight. Since the class sizes are imbalanced in the output variable (Table 1), we used weighted average of MSE (AMSE) and weighted average accuracy (AACC) across all classes (20).

Another criterion measures the association between true ($y$) and predicted ($\hat{y}$) labels using a rank order correlation statistic called Kendall’s correlation coefficient ($\tau_b$) (21). Based on this criterion, values of $\tau_b$ are in the interval of [-1,1]. Larger values of $\tau_b$ indicate better association between two ranking vectors (predicted and true output values) and hence better prediction. However, this measurement has a drawback since it does not consider the predictions individually and only takes into account the ranking of the prediction and true values as two vectors.

To tune the model hyperparameters, we used a Bayesian Optimization (BO) method presented in (22) to find the best set of hyperparameters for the model. BO is a strategy for optimizing black-box objective functions that are expensive to calculate. To evaluate the performance of the models in each iteration of BO, a 10-fold cross validation is implemented which splits the data to training, testing, and validation sets. The proportion of different GOSE levels was preserved among all training, testing and validation sets. To assess the performance of models, the same 10-fold cross validation approach was utilized, and the metrics were evaluated on the testing data.

In order to interpret the proposed model, we utilized Shapley Additive Explanation (SHAP), a unified framework for interpreting predictions via feature importance (23). SHAP unifies methods like LIME (24) and DeepLIFT (25) under the additive feature attribution umbrella. An additive feature attribution method formulates the outcome of a prediction model in the form of $f(x) = \theta_0 + \sum_{i=1}^{n} \theta_i x_{i}'$, in which $f$ is the prediction model, $\theta_i$ is the attribution assigned to each feature and $x_{i}'$ is a simplified input showing whether the $i$-th feature is missing. To implement the method, we used the SHAP python package written by the authors of the original paper.

**Results**

A total of 902 participants and 110 variables met the inclusion criteria for this study. Our proposed prediction model was trained on the training data and its performance was measured on the test set based on different metrics. The results are shown in Table 2. The values represent the mean and the standard error of mean for a 10-fold cross validation.
Table 2. Performance metrics for each model

<table>
<thead>
<tr>
<th>GCS range</th>
<th>Type of outcome</th>
<th>Models</th>
<th>AMSE</th>
<th>AACC</th>
<th>Kendall</th>
<th>AUC</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>All GCS scores</td>
<td>8-level GOSE</td>
<td>RNN</td>
<td>1.63 ± 0.11</td>
<td>0.24 ± 0.02</td>
<td>0.55 ± 0.04</td>
<td>0.59 ± 0.01</td>
<td>0.24 ± 0.02</td>
</tr>
<tr>
<td></td>
<td></td>
<td>IMPACT</td>
<td>2.02 ± 0.09</td>
<td>0.16 ± 0.04</td>
<td>0.38 ± 0.04</td>
<td>0.51 ± 0.03</td>
<td>0.16 ± 0.03</td>
</tr>
<tr>
<td>Binary outcome</td>
<td></td>
<td>RNN</td>
<td>0.35 ± 0.04</td>
<td>0.86 ± 0.03</td>
<td>0.75 ± 0.05</td>
<td>0.86 ± 0.03</td>
<td>0.91 ± 0.01</td>
</tr>
<tr>
<td>GCS ≤ 12</td>
<td>8-level GOSE</td>
<td>RNN</td>
<td>1.63 ± 0.13</td>
<td>0.21 ± 0.03</td>
<td>0.60 ± 0.04</td>
<td>0.58 ± 0.02</td>
<td>0.23 ± 0.03</td>
</tr>
<tr>
<td></td>
<td></td>
<td>IMPACT</td>
<td>2.08 ± 0.13</td>
<td>0.15 ± 0.03</td>
<td>0.38 ± 0.06</td>
<td>0.51 ± 0.03</td>
<td>0.14 ± 0.03</td>
</tr>
<tr>
<td>Binary outcome</td>
<td></td>
<td>RNN</td>
<td>0.42 ± 0.03</td>
<td>0.81 ± 0.02</td>
<td>0.64 ± 0.05</td>
<td>0.81 ± 0.02</td>
<td>0.82 ± 0.02</td>
</tr>
<tr>
<td></td>
<td></td>
<td>IMPACT</td>
<td>0.57 ± 0.04</td>
<td>0.66 ± 0.05</td>
<td>0.33 ± 0.10</td>
<td>0.66 ± 0.05</td>
<td>0.66 ± 0.05</td>
</tr>
</tbody>
</table>

We also analyzed the importance of time series data in the prediction task. To do so, we first only added the static data to the model and eliminated the recurrent part of the model (i.e., GRU-D units). We then trained another model which only incorporated the time series data. This model lacked the static inputs of the main model, so it only had the GRU-D units fed by time series data. The results of this analysis are shown in Table 3.

Table 3. Performance metrics for different models with only time series and static data

<table>
<thead>
<tr>
<th>Data used</th>
<th>AMSE</th>
<th>AACC</th>
<th>Kendall</th>
<th>AUC</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>All data</td>
<td>1.63 ± 0.11</td>
<td>0.24 ± 0.02</td>
<td>0.55 ± 0.04</td>
<td>0.59 ± 0.01</td>
<td>0.24 ± 0.02</td>
</tr>
<tr>
<td>Time series data</td>
<td>1.65 ± 0.13</td>
<td>0.25 ± 0.02</td>
<td>0.55 ± 0.03</td>
<td>0.56 ± 0.01</td>
<td>0.23 ± 0.01</td>
</tr>
<tr>
<td>Static data</td>
<td>2.66 ± 0.09</td>
<td>0.13 ± 0.01</td>
<td>0.15 ± 0.02</td>
<td>0.50 ± 0.01</td>
<td>0.07 ± 0.01</td>
</tr>
</tbody>
</table>

Figure 4 illustrates the important features that contributed to the outcome prediction of two sample patients with different outcomes. The important features were derived based on the magnitude of the SHAP values. Figure 5 shows the top important features derived using SHAP values. To assign a single value to each temporal feature, we took an average over all data for the first 120 hours of each patient. Figure 6 illustrates the time series values of three numerical variables among top features for favorable and unfavorable outcomes over the first five days of ICU stay. The missing values in the time series data were imputed using the linear interpolation.

Figure 4. Important features contributing to the outcome prediction of two patients. Patient A had a favorable outcome with GOSE eight, and patient B died after six months (i.e., GOSE equals one). Features indicated with red and blue colors contribute to the unfavorable and favorable outcomes, respectively. Patient A is a 24 years old male with both reactive pupils at ED, while patient B is a 42 years old female with only one reactive pupil at ED.
Figure 5. Top important features based on the absolute value of SHAP. Each point represents a patient, and the horizontal axis indicates the SHAP value of each feature for a patient. Negative and positive SHAP values imply the contribution to an unfavorable and favorable outcome, respectively. The color of each point represents the value of a feature for a patient. Feature names that start with “Daily” correspond to time series variables.

Figure 6. Time series values for different lab measurements over the first five days of ICU stay. Solid lines show the average and dashed lines are three standard errors. Favorable and unfavorable outcomes are separable for all three measurements.

Discussion

The main objective of this study is to predict 6-month outcome for the TBI patients admitted into the ICU. Prior studies using the TRACK-TBI dataset developed a clustering approach for identifying TBI cohorts (26) and implemented a linear regression model to predict the post-concussive symptoms among mild TBI patients (27).

Comparison of model performance and relevance

The comparison between the RNN model and IMPACT shows that the RNN model with time series data performs better based on all metrics. When all population is included and the outcome consists of all eight levels of GOSE, the AUC for both models is low (less than 0.6). One of the possible reasons for not achieving a high AUC is the implementation of a multiclass classification with eight classes. After dichotomizing the GOSE and refitting the models, the AUC for both IMPACT and RNN increases (0.69 and 0.86, respectively). However, RNN still outperforms the IMPACT on all metrics. Even though we did not include the CRASH model in our analysis, one study validated CRASH on TRACK-TBI dataset and demonstrated a poor discriminative ability (AUC of 0.49-0.50) for mild TBI patients (28).

Since the IMPACT model was originally developed on severe and moderate TBI patients, we expected a better performance for IMPACT on this population. However, compared to their performance on all populations, the RNN and IMPACT have a weaker performance. The RNN model with the binary outcome has an AUC of 0.86 on the whole cohort whereas it shows a lower AUC (0.81) on the severe and moderate TBI population. One potential reason is that TBI patients with lower acuity...
represented over half the studied cohort (555 subjects), thus by excluding these patients, training data shrinks by a significant amount.

Results in Table 3 illustrate that the time series data play an important role in the RNN model’s performance. The model trained with only time series data significantly outperforms the model trained with static data. For example, the Kendall coefficient for the model with time series data is 0.55 while this metric for the model with static data is only 0.15. The model trained on only time series data is even superior to the IMPACT model as all the metrics show a better performance. This improvement implies that training our model only with time series data demonstrates a better predictive ability in comparison with the IMPACT model.

**Interpretability of deep learning models**

To support the interpretability of the RNN model used in the work, we used SHAP to rank the features based on their contribution to the outcome prediction. The important features identified by our model highlight the validity of our model development strategy. The model found that age was one of the most important variables, which is a well-recognized observation (29) and indeed is a core component of the IMPACT prediction model. Other variables from the IMPACT model, namely motor GCS and pupillary reactivity, are also featured. Interestingly, the time series of the GCS are also featured in our model which is novel. This likely reflects the evolution of the patients’ neurological exam over time and therefore contributes to prognosis.

The inclusion of vital sign data is noteworthy. The diastolic blood pressure is a known risk factor for long-term cardiovascular health but has not been described in traumatic brain injury populations. Heart rate on the other hand may reflect shock or dysautonomia which impacts patients after TBI and may contribute to prognosis as well. The inclusion of sodium levels in our model suggests that the therapeutic intensity of treatment for elevations in intracranial pressure, which includes administration of hypertonic saline solution, is reflective of prognosis – the sicker patient is treated more aggressively. Therefore, elevated sodium as a negative prognostic marker likely represents the reality that the patient was exhibiting signs of more substantial injury requiring more aggressive therapy. However, this hypothesis would need to be formally tested as patients with the most severe injuries may develop diabetes insipidus (DI), which results in increased sodium, but DI remains relatively uncommon. Despite the limitations, it is worth highlighting that the model uncovered new and/or hypothesis-generating findings by leveraging both static and time-series data within this cohort.

The SHAP contributions of the top features cohere with their measured values, which is clearly illustrated in Figure 5. Most of the time, a consistent trend exists between SHAP contribution and measured values of each variable. Age is one of the best examples of this consistency between feature values and their SHAP contribution. As shown in Figure 5, the SHAP values decrease when the patients’ age increases, which means a patient’s age contributes more to an unfavorable outcome for older people. Figure 4 shows the SHAP contributions for patients A and B with favorable and unfavorable outcomes, respectively. Based on the SHAP scores, ED pupil reactivity contributes to the favorable and unfavorable outcomes for patients A and B, respectively, which matches the value of this feature since both pupils are reactive for patient A while only one of the pupils of patient B is reactive.

**Limitations**

One of the challenges in working with clinical time series data is that variables are not measured consistently over time. For example, vitals are recorded regularly, but mostly when patients are unstable. On the other hand, lab results are recorded only when physicians or nurses order them. As a result, clinical variables are recorded irregularly, and the measurement frequency varies between patients and is dependent on the place where the variables are taken. This frequency might be different across variables and even over time (16). However, our deep learning algorithm requires time series data to have regular time intervals. To make it possible, we discretize the observation time window into fixed-length time intervals (one-hour intervals) and aggregate all data within each interval. This method is a tradeoff between losing some information and increasing missing data. By increasing the length of time intervals, some information is lost since all available data in each interval is aggregated. Comparably, by decreasing the length of time intervals, as much data as possible is retained, but the missing data would increase since some of the intervals have no data.

Another limitation of this work is related to the amount of available data as well as external data for validation. The two notable prognostic models that are developed on large cohorts are IMPACT and CRASH. While both models are trained on data with more than 9,000 subjects, our model is trained only on 902 patients. Since the available data for this study is significantly lower than other major studies in the literature, providing more data would be beneficial for this prediction model. Furthermore, due to the lack of external data with enough time series data, we did not validate our model on an external dataset. We also acknowledge biases that may exist in the underlying data. For instance, if elevations in intracranial pressure were treated by a new medication that did not iatrogenically elevate the serum sodium concentrations, it might be expected that this laboratory value would no longer contain important prognostic information. However, if the model we present were not revised to reflect this change in practice, it could result in erroneous predictions. Methods for the
reproducibility and validation of models have yet to be standardized for clinical data science but are crucial in prognostic prediction applications.

**Conclusion**

We propose a deep RNN based model for long-term prognosis of TBI, which is trained on both static and time series data and predicts the GOSE after six months of injury. The model handles the time series missing values and utilizes the information from missingness patterns in temporal features. In summary, our results show that training the model on time series data for TBI patients can be informative and boost the performance of the predictions. Even the model that is solely trained on time series data outperforms the well-known IMPACT prediction model. Top important features are derived from the RNN model, and their values show a separable trend for favorable versus unfavorable outcomes. This study shows the magnitude of information that can be derived from time series data to prognose TBI more accurately.

**Acknowledgment**

This material is based upon work supported by the National Science Foundation under grants #1838730 and #1838745. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

The authors acknowledge all TRACK-TBI Study Investigators for providing access to data used in this work.

**References**


12. Little R, Rubin D. Statistical analysis with missing data [Internet]. 2019 [cited 2021 Mar 10].

Semantic Search for Large Scale Clinical Ontologies

Duy-Hoa Ngo, Madonna Kemp, Donna Truran, Bevan Koopman, Alejandro Metke-Jimenez
The Australian E-Health Research Centre, CSIRO, Australia

Abstract Finding concepts in large clinical ontologies can be challenging when queries use different vocabularies. A search algorithm that overcomes this problem is useful in applications such as concept normalisation and ontology matching, where concepts can be referred to in different ways, using different synonyms. In this paper, we present a deep learning based approach to build a semantic search system for large clinical ontologies. We propose a Triplet-BERT model and a method that generates training data directly from the ontologies. The model is evaluated using five real benchmark data sets and the results show that our approach achieves high results on both free text to concept and concept to concept searching tasks, and outperforms all baseline methods.

Introduction

Data standardisation is an important and challenging goal in the field of medicine. One of the key elements required to achieve semantic interoperability between clinical systems is the availability of common ontologies that define the concepts in the domain. Currently, the most comprehensive clinical ontology available is SNOMED CT, which contains more than 340,000 concepts and has been widely adopted in Electronic Health Record (EHR) systems worldwide. However, despite the availability and coverage of large ontologies such as SNOMED CT, there are still many challenges with adoption and implementation. These revolve around two main issues: 1) much of the source data is free text and needs to be mapped to a standard ontology; 2) existing systems use their own code systems and it is not feasible to replace them with SNOMED CT.

The first problem can be addressed by using natural language processing (NLP), whereby free text is analysed to identify relevant concepts. This process is called information extraction and it has been studied extensively in the biomedical domain. The process is typically divided in two phases: identifying the spans of text that represent relevant concepts and mapping these spans to concepts in a chosen ontology (also referred to as concept normalisation).

The second problem can be addressed by mapping between ontologies. For any ontology of considerable size these maps cannot be generated manually and require at least partial automation. This area of research is called ontology matching. Some of the most important elements that inform the matching process are the labels, synonyms and descriptions of the concepts.

Tailored solutions exist for both the concept normalisation problem and the ontology matching problem. However, these tend to be very specific to the particular use case. Instead, in this paper we propose a generalised method for searching large ontologies such as SNOMED CT, using short spans of text as input. This method can be used generically for both concept normalisation and ontology matching. A novel algorithm is presented, based on deep-learning-derived word representations, that is capable of finding good candidate concepts, even in the absence of common vocabulary. Input to the algorithm can be free text, when doing concept normalisation, or a concept from a source ontology, when doing ontology matching.

We empirically evaluate our method in a number settings — both concept normalisation and ontology matching. The results show that our method can effectively find relevant concepts, outperforming a number of comparison baselines. In addition, we show that our method is particularly suited to finding concepts where the input shares little or no common terms with the relevant concept.

Related Work

The problem of searching large ontologies has been studied in the context of data entry. Sevenster et al. proposed and evaluated an autocompletion algorithm for large medical ontologies and showed that a multi-prefix matching approach performs better than the baseline approach that only completes the entered string to the right. A modified version of this algorithm is implemented in Ontoserver, a high-performance FHIR terminology server, and it is used by default to do value set expansions. However, this algorithm doesn’t perform well on other tasks where the input strings are not partial prefixes but rather full words or short sentences. Also, the algorithm uses standard string matching so it only works well when the queries use the same vocabulary as the ontology being searched.

1 In FHIR, this is the operation used to implement auto-complete style widgets for data capture.
Searching ontologies has also been studied in the area of information extraction, specifically in the concept normalisation step where a span of text that has been identified as being relevant is mapped to a concept in an ontology. Wang et al.\textsuperscript{3} wrote an extensive literature review on clinical information extraction. An example of a state of the art algorithm specifically designed for the concept normalisation step can be found in the work of Luo et al.\textsuperscript{4}.

Finally, although not specific to ontology search, there is also relevant related work in the area of information retrieval (IR). Recently, advanced neural network methods have been developed to learn semantic representations of words and overcome the vocabulary mismatch problem of traditional IR models. Popular models that follow this trend include Word2Vec\textsuperscript{19}, GloVe\textsuperscript{22} and fastText\textsuperscript{21}. Their underlying idea is based on the distributional hypothesis in linguistics, i.e., words that are used and occur in the same contexts tend to purport similar meanings\textsuperscript{2}. Those neural networks are trained with large data resources by supervised learning algorithms. Once the training is finalised, every word located in the model’s dictionary will be encoded by a fixed length embedding vector. Then, those embedding vectors can be used as inputs to a combination function (e.g., average function) or another neural network model to derive an embedding vector of a query or a document. A limitation of these methods is that once the training is completed, the embedding vectors are static, which means that an embedding vector of a given word is always the same regardless of the context of use. Therefore, they may face issues with polysemy when a word might have a different meaning in a specific context.

Several contextualised word embedding methods based on deep long short-term memory (LSTM) architectures, such as CoVe\textsuperscript{24}, ELMO\textsuperscript{23} and FLAIR\textsuperscript{25}, have been proposed to improve the understanding of words and sentences. The main difference with the static word embedding methods is that the words’ embedding vectors are dynamically generated according to which context they have been used, i.e., surrounding words in a given sequence. Recently, transformer-based approaches like BERT\textsuperscript{17}, XLNet\textsuperscript{27}, RoBERTa\textsuperscript{20} have been proposed and achieved state-of-the-art results over most NLP downstream tasks. A key idea in these methods is that the meaning of a word in a sequence is represented by how much attention of that word attracts the other words in the sequence. Once the training of those models completes, they output a list of embedding vectors for all tokenised words of a given sequence. Then, those embedding vectors can be combined in different strategies to derive a semantic embedding vector for an input sequence.

A special feature of a search engine designed to search large ontologies is that the to-be-searched documents are concept labels, which are usually short and, therefore, sentence embedding methods are highly relevant because they can be used to compute semantic relatedness between this type of label. Recently, many sentence embedding methods, for example Doc2Vec\textsuperscript{20}, Skip-Thought\textsuperscript{10}, InferSent\textsuperscript{11}, Universal Sentence Encoder\textsuperscript{12} and the Sentence-BERT model\textsuperscript{8} have been proposed and have achieved good results in various natural language understanding tasks such as sentiment analysis, text classification, question answering and semantic textual similarity. Doc2Vec is an extension of Word2Vec that is trained with large, unlabelled text data. Skip-Thought is another extension of the Skip gram Word2Vec model that tries to predict the surrounding sentences of a given sentence. Universal Sentence Encoder trains a transformer network which augments unsupervised learning whereas InferSent trains a Siamese BiLSTM network with a max-pooling layer on top. Similar to the InferSent architecture, Sentence-BERT replaces a BiLSTM network with a BERT network and outperforms the other state-of-the-art methods on common semantic textual similarity and transfer learning tasks. These models have been trained on natural language inference data sets\textsuperscript{14, 15}, in which, an input to a learning model is a pair of sentences and the output is an inferred relation between them.

The algorithm we propose in this paper is most similar to Sentence-BERT\textsuperscript{8}. Sentence-BERT uses a Siamese architecture, and classification and regression objective functions. Our approach instead, uses a Triplet network\textsuperscript{9}, which processes three inputs in parallel, and a triplet loss function, which is a learning to rank metric for the three inputs.

**Method**

In this section we provide the details of our semantic search engine model for large scale clinical ontologies. First, we give an overview of the model and its components, and explain how to rank results for a given query. Then, we outline our training procedure, optimization objective in developing the model.

The main idea of this model is to transform every concept’s label into appropriate embedding vectors in a vector space so their locations preserve the semantic relations between concepts in the ontology. Figure 1 shows an example that illustrates this idea. The example shows that the concept Asthenia has three synonyms: “Weakness - general”, “Lassitude” and “Debility”. Due to the characteristics of synonymy, we would expect that the distance between the

\textsuperscript{2}https://en.wikipedia.org/wiki/Distributional_semantics
embedding vectors of synonyms, e.g., “Asthenia” vs. “Weakness - general”, would be smaller than the distance between the vectors of labels of concepts that are not synonyms, e.g., “Asthenia” vs. “Fatigue” or “Exhaustion”.

Figure 1: A SNOMED CT fragment for Asthenia, Feeling tired, Fatigue, Exhaustion and Energy and stamina finding. On the other hand, the concept Feeling tired is a sibling of the concept Asthenia because they both are children of the concept Fatigue. By applying the distance calculation method on the tree structure, the distance between two concepts is computed by the sum of the distances from those concepts to their lowest common ancestor. Therefore, we would also expect the distance between embedding vectors of a concept’s label to its direct parent concept’s label to be smaller than the distance of that concept’s label to its sibling concept’s label. In this example, the distance between concept Asthenia and concept Fatigue must be smaller than the distance between concept Asthenia and its sibling Feeling tired. Again, because “Lassitude” is a synonym label of Asthenia and “Weariness” is a synonym label of Fatigue, we would infer that a embedding vector of “Lassitude” is located closer to a embedding vector of “Weariness” than to a embedding vector of “Feeling tired”. The intuition of distance comparison based on synonymy and tree-based distance can be applied to all concepts in an ontology.

Label embedding with Triplet-BERT model

In order to achieve a vector space model that observes the properties described in the previous section, a Triplet-BERT model was trained to produce a semantic embedding vector for short text spans such as concepts' labels or user queries. The Triplet-BERT model is a kind of Triplet network and it was mentioned by Reimers et al. It consists of three instances of the same embedding layer containing a shared BERT network and a pooling layer (see Figure 2). It requires three text inputs, which are fed into the network at the same time, and represent three different roles: an anchor input, a positive input and a negative input. In this work, an anchor input is a user query (e.g., “Weakness - general”); a positive input is a label of a high relevant concept (e.g., “Asthenia”) to the user's query; and a negative input is a label of a less relevant concept (e.g., “Exhaustion”) to the query.

Figure 2: Triplet-BERT model architecture

The objective of a Triplet-BERT model is to train its parameters so at the end, the encoded embedding vector of the anchor input is closer to the embedding vector of the positive input than that to the embedding of the negative input. Firstly, each input is fed into a BERT network to produce a list of intermediate embedding vectors, then, a pooling layer combines these vectors to produce a summary embedding vector for a given input. Now, after receiving three embedding vectors $V_{anchor}$, $V_{positive}$ and $V_{negative}$ for anchor, positive and negative inputs respectively, the model computes distances between the embedding vector of the anchor input against the embedding vectors of positive and negative inputs. Finally, the Objective Function module compares the two distances and indicates whether the model needs to adjust its parameters through the back propagation algorithm. Once the training has completed, the embedding layer is used to produce embedding vectors for any user text query as well as concept labels in the ontology. In order to rank results for a given query, the cosine similarity metric is used to compute relevance scores between the user queries and the concepts’ labels.
Training Data Set

Deep neural network methods had been recently used to develop sentence embedding models. Due to the huge number of parameters in deep neural networks, these models require a significant amount of training data. Some common published data sets used for training sentence embedding are the Stanford Natural Language Inference (SNLI) data set\(^{14}\), the Multi-Genre NLI data set (MLNI)\(^{15}\) and the Semantic Textual Similarity (STS) data set\(^{16}\). Each entry in those data sets consists of a pair of sentences and a label that is either a relationship type or a semantic similarity score. These entries cannot be used in our Triplet network because it requires three input sentences at the same time. Therefore, we propose a method to generate a data set for training our Triplet network (see Algorithm 1) based on the aforementioned intuition about distances of concepts in an ontology’s hierarchy. The whole data set was generated from SNOMED CT and the Human Phenotype Ontology (HPO). Each entry in the data set consists of three strings following the same order: an anchor label, a positive label and a negative label. In total, the generated data set contains nearly 4 millions entries, which are then split into training, development and testing data sets with ratio of 90%, 5% and 5% respectively.

Algorithm 1: Generate training data from ontology for Triplet network

```
Input: T: Ontology
Output: D\text{train}, D\text{dev}, D\text{test}
D ← ∅
foreach concept ∈ T do
    conceptLabels ← getLabels(concept, T)
directParents ← getParents(concept, T)
otherConcepts ← getSiblings(concept, T) ∪ getSiblings(directParents, T)
foreach (label\(_1\) ≠ label\(_2\)) ∈ conceptLabels do
    parentLabel ← getRandomLabel(directParents, T)
    otherLabel ← getRandomLabel(otherConcepts, T)
    addToDataset(D, anchor=label\(_1\), positive=label\(_2\), negative=parentLabel)
    addToDataset(D, anchor=label\(_1\), positive=label\(_2\), negative=otherLabel)
    addToDataset(D, anchor=label\(_1\), positive=parentLabel, negative=otherLabel)
end
d\text{train}, d\text{dev}, d\text{test} ← splitTrainDevTest(D)
```

Training Details

**BERT network.** Transfer learning was used to fine-tune BERT parameters. In this work, we adopted BioBert-Base v1.1\(^{13}\) — a state-of-the-art biomedical language representation model, which has been widely using biomedical natural language processing tasks.

**Pooling layer.** The pooling layer is added on top of the BERT network to get an embedding vector for a given text input. Different strategies can be used to work with BERT’s output embedding vectors, however, according to Reimers et al\(^8\), the MEAN strategy achieved a better result than the others. Therefore, we chose the MEAN strategy for the pooling layer to produce a fixed size, i.e., a 768-dimensional embedding vector for the given text input.

**Distance metric.** We use Euclidean distance to compute distances between the anchor embedding vectors \(V_{anchor}\), the positive embedding vectors \(V_{positive}\) and the negative embedding vectors \(V_{negative}\).

**Objective function.** The objective of our model is to move the anchor embedding vector \(V_{anchor}\) closer to the positive embedding vector \(V_{positive}\) and far away from the negative embedding vector \(V_{negative}\). Therefore, we minimize the following objective function to tune the model’s parameters:

\[
loss = \max(||V_{anchor} - V_{positive}|| - ||V_{anchor} - V_{negative}|| + m, 0)
\]

Here, a small margin value \(m\) is used to push the distance \(||V_{anchor} - V_{negative}||\) being at least \(m\) higher than the distance \(||V_{anchor} - V_{positive}||\). Otherwise, the loss value is positive, thus its derivatives to model’ parameters are not 0, so the back propagation algorithm will update the model’s parameters. In the training phase, we fix \(m = 0.1\).
**Training settings.** Our model was trained in 5 epochs. We set a batch-size of 32, Adam optimizer with learning rate 2e-5, and a linear learning rate warm-up over 10% of the training data. The training process was done in 40 hours with one GPU with 32G RAM, using Python 3.6, Pytorch 1.6 and CUDA 10.1. The Triplet-BERT codes was derived from Sentence-BERT codes by replacing siamese network by triplet network.

**Evaluation**

In this section, we firstly describe how to evaluate the performance of our semantic search system and related evaluation metrics. Next, we present data sets used in the evaluation and finally we analyze the experimental results.

As our evaluation measure we use Hits@K. For a given query, Hits@K is 1 if the relevant concept is found in the top K results; otherwise it is 0. In our evaluation, we used Hits@1, Hits@5 and Hits@10. Additionally, in order to measure the usefulness of the list of returned results for a given query, Normalized Discounted Cumulative Gain (nDCG) and Mean Reciprocal Rank (MRR) were used. Their underlying assumption is that the higher the relevant results are ranked, the more gain the user receives. Therefore, the ideal ranking would first return the result with the highest relevance level, then the next highest relevance level, etc. nDCG@K measures the performance of a search system based on the relevance order of the K returned results against the ideal ranking order. In our evaluation, we used nDCG@1, nDCG@5 and nDCG@10.

**Data Sets for Evaluation**

For evaluation, the following five data sets were used:

- **cadec2sct**: This data set contains 2036 unique short text extracted and annotated to SNOMED CT clinical finding concepts from medical forum posts on patient reported Adverse Drug Events (ADEs).
- **note2sct**: This data set contains 4960 unique short text extracted and annotated to SNOMED CT clinical finding concepts from real patients’ narrative discharge summaries in a hospital in Queensland, Australia.
- **hpo2sct**: This data set contains 14,149 unique labels collected from 5978 phenotype concepts from the Human Phenotype Ontology (HPO), which have been mapped to concepts in SNOMED CT.
- **fma2sct**: This data set contains 13,123 unique labels collected from 5702 concepts from the Foundational Model of Anatomy Ontology (FMA), which have been mapped to concepts in SNOMED CT.
- **ncit2sct**: This data set contains 46,185 labels collected from 13,830 concepts from the National Cancer Institute’s Thesaurus (NCIt) which have been mapped to concepts in SNOMED CT.

Three of the data sets, i.e., **cadec2sct**, **note2sct** and **hpo2sct** were manually annotated by two clinical terminology experts, whereas, the other two, i.e., **fma2sct** and **ncit2sct** were taken from the Large BioMed Track in the Ontology Alignment Evaluation Initiative 2020.

The first two data sets aim to evaluate the text-to-concept searching functionality where a user runs free-text queries to search relevant SNOMED CT concepts. In the **cadec2sct** data set, clinical symptoms or diseases were described by different members with or without medical background, so the texts were not restricted to follow any standard naming rules. The entered texts were just the observation or understanding of lay people reporting adverse drug events. In contrast, in the **note2sct**, clinical entities were entered by doctors in hospital. Those texts are in free-text form, but the vocabulary is expected to be technical in nature.

The last three data sets were extracted from well-designed biomedical ontologies. Thanks to cross-ontology alignment, it is expected that the results of searching a concept’s label from one ontology will return its mapping concepts from another ontology that is found in the alignment. Additionally, these data sets can also be used to evaluate SNOMED CT concept-to-concept search when all labels of a given concept from a source ontology, such as HPO, FMA or NCIt, are used to search relevant SNOMED CT concepts.

**Models for Comparison**

As presented in section Method, the key operation in our work is how to encode a short text into an embedding vector. For evaluation and comparison purposes, the following baseline embedding methods have been implemented to build

---

3http://www.cs.ox.ac.uk/isc/projects/SEALS/oaei/

4http://oaei.ontologymatching.org/
different semantic search systems for SNOMED CT.

- **Elasticsearch BM25**: BM25 defines a weight for each term as a product of some IDF-function and some TF-function and then summarises that term weight as the score for the whole document towards the given query. In this work, a document is a collection of labels of a SNOMED CT concept.
- **Word2Vec-based Average**: This model defines an embedding vector of a given text as an average of embedding vectors, which map to text’s tokens in a pre-trained biomedical word2vec resource⁵.
- **BioBERT based CLS**: This model uses a pre-trained BioBERT⁶ to encode an input text into a sequence of corresponding embedding vectors. Then, it defines an embedding vector of a given text by embedding a vector of special tokens [CLS] according to the original idea from BERT⁷.
- **BioBERT based MEAN**: Similar to the previous model but this model computes an average of all output embedding vectors as a summarised embedding vector for a given input.
- **Triplet-BERT**: We use the fine-tuned BioBERT with MEAN strategy for the pooling layer to encode a given text into an embedding vector.

The first model is a keyword-based search engine, which does not count synonymous features in its scoring function. The second model uses a pre-trained word embedding, trained on a very large biomedical text collection. Each word in a pre-trained word2vec always has the same embedding vector regardless of the context, so this model can be considered as a context-free embedding model. The third and the fourth models are based on a pre-trained BERT network, in which a word may have different embedding vectors depending on the context of use. The two models can be considered as general contextualised embedding models. The last model fine-tunes BERT parameters over the triplet data set generated from SNOMED CT, and therefore it can be considered as a domain-specific contextualised embedding model.

**Evaluation on Concept Normalisation Task**

In this experiment, a user provides a short query string and asks the system to return relevant concepts from SNOMED CT. The query is either a clinical mention in cadec2sct and note2sct or a concept’s label in HPO, FMA and NCIt. We split the five data sets above into two groups based on the different editions of SNOMED CT used in the ground truth data. The first group consists of three data sets including cadec2sct, note2sct and hpo2sct, where the ground truth data uses clinical finding concepts from the Australian edition of SNOMED CT. The second group consists of two data sets, nci2sct and fma2sct, where the ground truth was created from concepts in the international edition of SNOMED CT.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>cadec2sct</th>
<th>note2sct</th>
<th>hpo2sct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size</td>
<td>#queries=2,036</td>
<td>#queries=4,960</td>
<td>#queries=14,149</td>
</tr>
<tr>
<td>K</td>
<td>K=1</td>
<td>K=5</td>
<td>K=10</td>
</tr>
<tr>
<td>BM25</td>
<td>0.132</td>
<td>0.240</td>
<td>0.307</td>
</tr>
<tr>
<td>Word2Vec</td>
<td>0.191</td>
<td>0.359</td>
<td>0.418</td>
</tr>
<tr>
<td>BERT-CLS</td>
<td>0.114</td>
<td>0.216</td>
<td>0.283</td>
</tr>
<tr>
<td>BERT-MEAN</td>
<td>0.136</td>
<td>0.266</td>
<td>0.319</td>
</tr>
<tr>
<td>Triplet-BERT</td>
<td><strong>0.385</strong></td>
<td><strong>0.603</strong></td>
<td><strong>0.654</strong></td>
</tr>
</tbody>
</table>

**Table 1**: Hits@K evaluation of searching a query to SNOMED CT clinical finding concepts.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>fma2sct</th>
<th>nci2sct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size</td>
<td>#queries=13,123</td>
<td>#queries=46,185</td>
</tr>
<tr>
<td>K</td>
<td>K=1</td>
<td>K=5</td>
</tr>
<tr>
<td>BM25</td>
<td>0.310</td>
<td>0.571</td>
</tr>
<tr>
<td>Word2Vec</td>
<td>0.183</td>
<td>0.506</td>
</tr>
<tr>
<td>BERT-CLS</td>
<td>0.157</td>
<td>0.244</td>
</tr>
<tr>
<td>BERT-MEAN</td>
<td>0.226</td>
<td>0.448</td>
</tr>
<tr>
<td>Triplet-BERT</td>
<td><strong>0.700</strong></td>
<td><strong>0.855</strong></td>
</tr>
</tbody>
</table>

**Table 2**: Hits@K evaluation of searching a query to all SNOMED CT concepts

Table 1 and Table 2 show the results computing Hits@K values in the two groups with K = 1, 5 and 10 for the five aforementioned methods. Our method Triplet-BERT consistently outperforms all other methods on all metrics over

---

⁵https://bio.nlplab.org/
⁶https://github.com/dmis-lab/biobert
all five data sets. We have also computed the statistical significance \textbf{p-value} for Triplet-BERT on a paired \textit{t-test} against the remaining methods. In almost all cases with various datasets and ranking value K, the computed value \( p \ll 0.05 \), which indicates strong evidence of significant differences between the results obtained by Triplet-BERT and the results returned from BM25, Word2Vec, BERT-CLS and BERT-MEAN. In Table 1 and Table 2, we highlighted cases where \( p > 0.05 \) by underlining the corresponding cell. For example, for dataset \texttt{note2sct} and \( K = 5 \), the \textbf{p-value} computed for Triplet-BERT and Word2Vec is 0.75.

On the other hand, Triplet-BERT achieved a high Hits@10 value 0.904 for data set \texttt{note2sct}, which means for a clinical mention written in the narrative of a discharge summary, the chance is about 90% that annotators can find the correct SNOMED CT concept within the top ten results that are returned. Similarly, the Hits@10 values are high for the \texttt{hpo2sct} and \texttt{fma2sct} data sets. Good results on these data sets were expected as the text in \texttt{note2sct}, \texttt{hpo2sct} and \texttt{fma2sct} data sets mainly focus on clinical finding concepts; the \texttt{fma2sct} data set mainly focuses on anatomical structure concepts, and both types of concepts are comprehensively covered by SNOMED CT.

The performance of Triplet_BERT on \texttt{cadec2sct} and \texttt{nci2sct} datasets is a bit lower than on the \texttt{note2sct}, \texttt{hpo2sct} and \texttt{fma2sct} data sets. This is likely because of the quality of the text in \texttt{cadec2sct}. Many references were written in casual language which leads to irrelevant concepts being retrieved. For example, “threw up” was annotated to 422400008 | Vomiting (disorder) |, but our system returned 282667008 | Does throw (finding) |. Additionally, the lack of context also causes ambiguity; for example, the text “damage to my muscles” was annotated to 129565002 | Myopathy (disorder) |, which means a disorder of skeletal and/or smooth muscle, but our system returned 95847005 | Injury of muscle (disorder) |, which is a child concept of the annotated concept. The second reason for lower performance was posited as lower quality of synonymous labels of NCIt concepts in the \texttt{nci2sct} data set. For example, an NCIt concept \texttt{C1212} has the following labels: “sirolimus”, “rapamycin”, “SILA 9268a”, “WY-090217”, “AY 22989”, “rapamune” and “rapa”, but only the first two labels were found by our search system. The other labels contain either numeric tokens or abbreviations that do not provide helpful information to find relevant concepts.

An interesting observation here is that without the fine-tuning of Triple_BERT, the basic BERT models did not out-perform Elastic BM25 and Word2Vec (even though one may expect them too). An explanation for this phenomenon is that the Elastic BM25 and Word2Vec mainly focus on keywords similarity, in which all stop-words had been removed from the text during indexing and searching. On the contrary, the original BERT model creates embedding vectors for all tokens of the text, including stop-words. In the case of BERT, due to the short query text and concept labels (≈1-2 tokens), BERT’s self-attention layers may not capture the context of the text’s tokens. For example, all the terms “Headache”, “head pain”, “Cephalodynia”, “Cephalalgia” and “Cephalgia” refer to the same meaning - “pain in head”. Pre-trained BioBERT computes similarity scores between (headache, cephalodynia) = 0.69; between (headache, cephalgia) = 0.73; and between (cephalodynia, cephalgia) = 0.97. The big difference in similarity scores tell us that in pre-train BioBERT, there is not enough context to embed those terms in highly similar vectors. This weakness is solved in Triplet_BERT because the model was fine-tuned from the original BERT model to push embedding vectors of those terms to be close to each other.

\textbf{Impact of Overlapping Text and Synonymy in Searching Performance}

In this experiment, we investigated the performance with respect to how similar queries are to their relevant concepts. Queries that are very similar to a relevant concept label will be easy to match, while queries that share no common terms will be harder to match. First, we define an overlapping degree as the proportion of shared tokens between query and concept label. Assuming that a query \( q \) contains a list of non stop-words: \( T_q = \{ t_{q_1}, \ldots, t_{q_N} \} \), and similarly, the concept \( c \) corresponding to the query \( q \) in the ground truth data sets contains a list of non stop-words: \( T_c = \{ t_{c_1}, \ldots, t_{c_M} \} \) of all its labels, then an overlapping degree of \( q \) against \( c \) is calculated as follows: \textit{overlapping}(q, c) = \frac{||T_q \cap T_c||}{||T_q||}. \) We divide our evaluation into two: 1) clinical findings concepts, which are often clearly expressed and have many synonym labels (\texttt{cadec2sct}, \texttt{note2sct} and \texttt{hpo2sct}); 2) all concepts types but with fewer synonym labels (\texttt{fma2sct} and \texttt{nci2sct}). We do this split to understand the impact that the synonym labels have for the different methods.

Figure 3 shows the line charts of Hits@10 values of different methods at different intervals of overlapping degrees on the two aforementioned groups. (The left hand side of the plots shows queries with many shared terms; the right
Evaluation on only clinical findings concepts with more synonym labels (cadec+note+hpo).

Evaluation on only clinical findings concepts with fewer synonym labels (fma+nci).

**Figure 3:** Evaluation of searching performance with different overlapping degrees

hand side shows queries with little or no term overlap.) As the overlapping degree goes down, the performance of all methods also decreases because these are harder queries to match. The line charts show that the Elastic BM25 method achieved high performance when the queries share many terms with their corresponding concepts. Particularly, in the range of overlapping from 0.8 – 1.0, the Hits@10 values of BM25 method was 0.943 for the cadec+note+hpo group, and 0.977 for the fma+nci group. Those values were close to the Hits@10 values of our Triplet-BERT method, i.e., 0.982 and 0.990 respectively. However, when the overlapping degree gradually decreases, the Hits@10 values of the BM25 method quickly falls to 0.0. In contrast, Triplet-BERT still maintains its performance as the overlap decreases. The reason for this is that Triplet-BERT encodes the meaning of terms; thus it is still able to retrieve the relevant concept for a query even when it shares little or no common terms. Comparison between Figure 3a and Figure 3b, shows that number of synonym labels does not have a dramatic effect on performance for all five methods considered. A point to note, though, is that when Triplet-BERT has access to queries with more synonyms (Figure 3a), it maintains better performance for very low overlap queries compared with fewer synonyms (Figure 3b). This shows that Triplet-BERT does exploit synonyms for better performance.

Let us consider some specific examples of matching with low overlap queries. A query “narrow retinal arterioles” should match 271728000 | Retinal arteries attenuated (finding). After running this query with five methods described above, we found that Elastic BM25 failed to return its correct result, whereas other methods were successful. An explanation is that the query and its corresponding concept shared only one word, “retinal”, which is 1/4 of the query length. On the other hand, because “attenuated” vs. “narrow” as well as “arteries” vs. “arterioles” are semantically similar, the other methods, which rely on the meaning of the words, are able to find the correct result. Another interesting example is the query “tooth mass excess”, which matches SNOMED CT concept 71485000 | Macroodontia (disorder). The overlap degree is 0.0. After running this query, only Triplet-BERT found the correct result.

**Evaluation on the usefulness of the searching results**

Search system are often evaluated according to the Normalized Discounted Cumulative Gain (nDCG) metric. We do the same here but first need to define the respective gain function a user receives for different types of results. Assume a user runs a query $q$. The ‘gain’ $g$ a user receives for a result list $l$ against the correct result $q^*$ is defined as follows: $g = 3$ if $l = q^*$; $g = 2$ if $l$ is a direct parent or direct child of $q^*$ in the ontology; $g = 1$ if $l$ is a grand parent, a grand child, a uncle or a sibling of $q^*$ and $g = 0$ otherwise.

Table 3 shows nDCG@1, nDCG@5, nDCG@10 and MRR values of different methods on two group datasets described in the previous section. By all metrics, Triplet-BERT method achieved the best performance. The average nDCG values produced by the Triplet-BERT is around 70%, which means the order of the returned results is highly correlated to the order of ideal results.

Now, let’s see an example to illustrate the importance of nDCG and MRR values. Assume that a user wants to find relevant SNOMED CT concepts for the query “delayed closure of fontanels”. Our Triplet-BERT method returned a
Table 3: nDCG@K and Mean Reciprocal Rank (MRR) evaluation of searching methods

<table>
<thead>
<tr>
<th>Datasets</th>
<th>More synonyms (cadec+note+hpo)</th>
<th>Less synonyms (fma+nci)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>nDCG@K</td>
<td>MRR</td>
</tr>
<tr>
<td></td>
<td>K=1</td>
<td>K=5</td>
</tr>
<tr>
<td>BM25</td>
<td>0.587</td>
<td>0.644</td>
</tr>
<tr>
<td>Word2Vec</td>
<td>0.632</td>
<td>0.687</td>
</tr>
<tr>
<td>BERT-CLS</td>
<td>0.491</td>
<td>0.562</td>
</tr>
<tr>
<td>BERT-MEAN</td>
<td>0.520</td>
<td>0.596</td>
</tr>
<tr>
<td>Triplet-BERT</td>
<td><strong>0.758</strong></td>
<td><strong>0.799</strong></td>
</tr>
</tbody>
</table>

Correct concept 82779003 | Late fontanel closure | in the first rank; the Word2Vec method returned this concept in second rank; both BERT-CLS and BERT-MEAN methods returned it in fifth position, whereas Elastic BM25 returned it in seventh position. That means that the user immediately finds the correct answer at the first or the second look if the system is based on Triplet-BERT or Word2Vec. In contrary, the user must spend more time to trace along the list of results to get a correct answer if the system uses BERT-CLS, BERT-MEAN or Elastic BM25. In this example, for Triplet-BERT: \( MRR = \frac{1}{1} = 1.0 \); for Word2Vec: \( MRR = \frac{1}{2} = 0.5 \); for BERT-CLS and BERT-MEAN: \( MRR = \frac{1}{5} = 0.2 \) and for Elastic BM25: \( MRR = \frac{1}{7} = 0.14 \). So, the higher value of \( MRR \) is, the less time the user spends to find the correct answer for a given query.

Doing further analysis on this example, Triplet-BERT returns top five results including: the correct concept in the first position, its sibling concept 1667003 in the second position, its parent concept 248382004 in the third position, an uncle concept 249079005 in the fourth position and a grandparent concept 248381006 in the fifth position. Based on the respective gain defined above, Triplet-BERT obtains an \( nDCG@5 = 0.976 \). This number shows that the order of the returned results is very close to the ideal order, which means that the user can find not only the correct concept but also its close neighbours.

Despite the fact that all five methods have found the correct results in their top 10 returned lists, which means the all have the same Hits@10 score, the order of returned results is different. This difference impacts the time that a user needs to find the correct answer. Our experiment shows that on average, the results obtained from Triplet-BERT were more useful than those from other baseline methods.

Evaluation on ontology matching task

In this experiment, we look for relevant concepts in an ontology for a given concept from another ontology. This is ontology matching task, where each concept from an ontology can be mapped to one or several concepts from the other ontology. The concept-to-concept search slightly differs from the previous text-to-concept search in the way a the query is formulated. In concept normalisation (text-to-concept), a single query for a short text or a label was executed, whereas in concept-to-concept search, multiple queries, i.e., labels of a query concept, will be run.

Table 4: Hits@K evaluation on concept to concept searching

<table>
<thead>
<tr>
<th>Dataset</th>
<th>hpo2sct</th>
<th>fma2sct</th>
<th>nci2sct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size</td>
<td>#concepts=5978</td>
<td>#concepts=5702</td>
<td>#concepts=13830</td>
</tr>
<tr>
<td>Hits@K</td>
<td>K=1</td>
<td>K=5</td>
<td>K=10</td>
</tr>
<tr>
<td>BM25</td>
<td>0.529</td>
<td>0.718</td>
<td>0.778</td>
</tr>
<tr>
<td>Word2Vec</td>
<td>0.595</td>
<td>0.781</td>
<td>0.834</td>
</tr>
<tr>
<td>BERT-CLS</td>
<td>0.459</td>
<td>0.627</td>
<td>0.684</td>
</tr>
<tr>
<td>BERT-MEAN</td>
<td>0.549</td>
<td>0.740</td>
<td>0.794</td>
</tr>
<tr>
<td>Triplet-BERT</td>
<td><strong>0.770</strong></td>
<td><strong>0.926</strong></td>
<td><strong>0.947</strong></td>
</tr>
</tbody>
</table>

Table 4 shows the Hits@1, Hits@5 and Hits@10 values for concept normalisation. Triplet-BERT outperforms other methods on all metrics. If Hits@10 is considered, there is a \( \approx 94\% \) chance that Triplet-BERT find a relevant SNOMED CT concept be found for a given HPO, FMA and NCI concept. A high recall at top 10 searching results can be used as an input to automate ontology matching\(^29\). It would also greatly reduce the time and cost on manual ontology alignment.
Conclusion

In this work, we proposed Triplet-BERT — a label embedding model that can be used to build a semantic search system for large scale clinical ontologies. We also proposed a method for generating a training data set for the model directly from an ontology. The method is generic in nature and can be used for both concept normalisation and ontology matching. Several experiments were conducted using SNOMED CT and showed that the proposed method outperforms baseline methods such as Elastic BM25, Word2Vec and BERT in all evaluation metrics on five benchmark data sets. In particular, the method was effective at mapping queries that had little or no common terms with relevant concepts. The strong empirical results suggest that Triplet-BERT can be used as the basis for both automatic ontology matching algorithms and searching tools to assist humans building ontology maps.

References

Towards a framework for comparing functionalities of multimorbidity clinical decision support: A literature-based feature set and benchmark cases

Dymphna O’Sullivan, PhD¹, William Van Woensel, PhD², Szymon Wilk, PhD³, Samson W. Tu, MS⁴, Wojtek Michalowski, PhD⁵, Samina Abidi, MD PhD⁶, Marc Carrier, MD⁷, Ruth Edry, MD⁸, Irit Hochberg, MD⁸, Stephen Kingwell, MD⁷, Alexandra Kogan¹⁰, MSc, Martin Michalowski, PhD¹¹, Hugh O’Sullivan, MD¹², Mor Peleg, PhD¹⁰

¹ASCNet Research Group, Technological University Dublin, Dublin, Ireland
²NICHE Research Group, Dalhousie University, Halifax, Canada
³Institute of Computing Science, Poznan University of Technology, Poznan, Poland
⁴Center for BioMedical Informatics Research, Stanford University, Stanford, CA, 94305, USA
⁵Telfer School of Management, University of Ottawa, Ottawa, ON, Canada
⁶Medical Informatics Faculty of Medicine, Dalhousie University, Canada
⁷The Ottawa Hospital, Ottawa, ON, Canada
⁸Bruce Rappaport Faculty of Medicine, Technion - Israel Institute of Technology, Haifa, Israel
⁹Rambam Medical Center, Haifa, Israel
¹⁰Department of Information Systems, University of Haifa, Haifa, Israel, 3498838
¹¹School of Nursing, University of Minnesota, Minneapolis, MN, USA
¹²BJD Family Practice, Ballyjamesduff, Cavan, Ireland

Abstract

Multimorbidity, the coexistence of two or more health conditions, has become more prevalent as mortality rates in many countries have declined and their populations have aged. Multimorbidity presents significant difficulties for Clinical Decision Support Systems (CDSS), particularly in cases where recommendations from relevant clinical guidelines offer conflicting advice. A number of research groups are developing computer-interpretable guideline (CIG) modeling formalisms that integrate recommendations from multiple Clinical Practice Guidelines (CPGs) for knowledge-based multimorbidity decision support. In this paper we describe work towards the development of a framework for comparing the different approaches to multimorbidity CIG-based clinical decision support (MGCDS). We present (1) a set of features for MGCDS, which were derived using a literature review and evaluated by physicians using a survey, and (2) a set of benchmarking case studies, which illustrate the clinical application of these features. This work represents the first necessary step in a broader research program aimed at the development of a benchmark framework that allows for standardized and comparable MGCDS evaluations, which will facilitate the assessment of functionalities of MGCDS, as well as highlight important gaps in the state-of-the-art. We also outline our future work on developing the framework, specifically, (3) a standard for reporting MGCDS solutions for the benchmark case studies, and (4) criteria for evaluating these MGCDS solutions. We plan to conduct a large-scale comparison study of existing MGCDS based on the comparative framework.

Introduction

Most recommendations from clinical practice guidelines (CPGs) focus on the management of single diseases. Such recommendations may be harmful or impractical for patients with multimorbidity. Multimorbidity has been defined as one of the “grand challenges in clinical decision support” by Sittig et al¹ because of a difficulty with creating mechanisms to identify and eliminate redundant, contraindicated, potentially discordant, or mutually exclusive guideline-based recommendations. Using a computer language such as PROforma or GLIF3, one can computerize CPG as Computer-interpretable guidelines (CIGs)²,3 that enable CIG-driven clinical decision support systems (CDSS). These have typically addressed a single morbidity as per the single-disease focus of the CPG they are ultimately based on. However, the rise in multimorbidity is driving the need for complex treatment plans with many potential interactions and adverse events among CIGs. Several researchers have begun developing multimorbidity CIG-based
clinical decision support (MGCDS) that can detect and mitigate interactions among recommendations belonging to different CIGs in order to develop non-conflicting management plans for patients with multimorbidity. Given the number of active research groups and different MGCDS approaches, we decided to create a standard framework for comparing these approaches. Our framework consists of (1) the relevant features of the multimorbidity problem that should be addressed, (2) a set of benchmarking case studies that are representative and cover the features, (3) a standard for reporting the MGCDS solutions for the benchmarks, and (4) criteria for evaluating these solutions. We intend for this framework to be used in standardized evaluations of prior or new/updated MGCDS, to help assess their functionalities, and identify gaps in the state-of-the-art. This paper reports results of research on the first two parts of the framework. We will use the complete benchmark framework in a comparative MGCDS study, which is currently underway and involves a number of groups who have previously developed MGCDS. A companion paper, to be published after the comparative study has completed, will report on the development of part 3 and part 4 of the framework as well as the comparative study results themselves.

**Background and related work**

MGCDS has gained attention in recent years and multiple approaches have been proposed and described in the literature. A comprehensive review and comparison of relevant approaches are provided elsewhere and here we summarize the relevant characteristics (e.g., an employed CIG representation formalism or reasoning method) of most recent proposals. We also point at clinical case studies that were used to test these approaches. We applied a semi-formal process to search for relevant publications. Specifically, we identified candidate papers based on the description of related works in two recent publications on MGCDS by Kogan et al. and Michalowski at al. Then we limited the list of candidates to journal papers published in the last 5 years. Finally, we screened the eligible papers to select those that present the latest versions of MGCDS methodologies and illustrate their applications in case studies.

Kogan et al. propose a goal-oriented MGCDS which frames patient management as a goal attainment problem. The methodology, implemented in the form of the GoCom system, relies on CIGs represented in PROforma and augmented with knowledge about goals and physiological effects of specific CIG tasks. This knowledge is encoded with standardized terms coming from controlled terminologies (e.g., SNOMED CT) and medical ontologies (e.g., NDF-RT). The planning algorithm used by GoCom operates on goal forests that capture goals associated with specific CIGs applied to the patient, identifies inconsistencies among goals and mitigates them by proposing alternative solutions. GoCom is able to reason at different abstraction levels (e.g., specific drugs and their classes) by exploring external ontologies and to generate explanations for specific solutions. Alternative solutions, together with supporting explanations, are then presented to the clinician who makes the final decision. GoCom relies on the HL7 FHIR standard to represent clinical data, thus facilitating integration with existing hospital information systems (HISs). GoCom was tested on 6 clinical scenarios involving concurrent application of multiple (2-3) CIGs -- one of the scenarios describing a patient being managed for stroke and duodenal ulcer (DU) and diagnosed with osteoporosis was used as the basis for development of Case 1 described in this paper. Moreover, GoCom was evaluated by medical students and interns in two empirical studies that confirmed its usefulness and validity of delivered recommendations.

Michalowski at al. describe MitPlan – a planning-oriented MGCDS that frames multimorbid patient management as a planning problem that allows to mitigate adverse interactions between multiple CIGs and to derive safe management plans. MitPlan constructs a plan for a given time horizon that optimizes some objective function (e.g., overall cost) – with time horizon and objective function being specified by the clinician. It also considers patient preferences – if several options are possible, it selects the most preferred one. Unlike GoCom, MitPlan generates a single optimal management plan that is presented to the clinician for approval. It accepts CIGs represented as Actionable Graphs (AGs) that are based on the task-network model and can be easily derived from other representations, such as GLIF or PROforma. AGs are automatically transformed to Planning Domain Definition Language (PDDL) for further processing. PDDL is also used to represent secondary domain knowledge on possible interactions and strategies to mitigate them. MitPlan was evaluated by collaborating physicians using a scenario describing a patient suffering from chronic kidney disease (CKD) and hypertension (HT), who experiences an acute episode of atrial fibrillation (AF). This scenario was used as a basis for development of Case 2 described in this paper.

Jafarpour et al. propose an ontology-based MGCDS for execution-time integration of multiple CIGs. The approach assumes integration points need to be first identified by the clinician, and then appropriate integration policies are instantiated and applied to mitigate adverse interactions at execution-time. Integration policies are defined using the CIG-IntO ontology that is represented in OWL and processed by a standard OWL reasoner (the authors employ Jena). The approach is able to discover drug-drug and drug-disease interactions using the Bio2RDF DrugBank ontology, handle temporal aspects of mitigation (e.g., delaying tasks to avoid conflicts), and rollback integration policies that...
are no longer safe or efficient at execution-time. CIG-IntO also allows for defining conditional integration policies where revisions introduced to CIGs may be further customized depending on additional conditions, e.g., related to the patient’s health profile. This increases the flexibility and generality of integration policies. Moreover, in addition to mitigation integration policies the authors also propose optimization integration policies that minimize resource use and costs by removing redundant tasks from management plans and re-using test results. The approach was tested using 6 case studies with pairs of CIGs and positively evaluated by a panel of health informaticians and physicians.

Another MGCDs is presented by Fdez-Olivares et al. who propose a Multi-Agent Planning (MAP) framework. The MAP framework relies on Hierarchical Task Networks (HTNs) to represent and control the planning process and involves multiple agents that develop different candidate management plans. Possible plans are evaluated using an objective function that considers plan cost and complexity assessed according to the patient’s quantitative preferences. Finally, the optimal plan is presented to the clinician for approval. MAP accepts CIGs represented in Hierarchical Planning Description Domain Language (HPDL) – such representation can be obtained from the CIGs modelled in Asbru formalism. Proposed approach uses HPDL to capture possible adverse interactions and patient’s qualitative preferences related for example to the mode of drug administration or frequency of administration – these preferences are considered when constructing candidate plans. As with the approach by Jafarpour et al., MAP takes into account temporal aspects of interactions and mitigation. MAP was evaluated using a case study involving a patient with diabetes mellitus (DM) and hypertension (HT) and managed at different time points of disease progression and treatment process.

Piovesan et al. describe the MGCDs implemented as the GLARE-SSCPM system. The proposed system relies on multiple methods to identify interactions and mitigate them, such as temporal reasoning, cost-benefit analysis and model-based verification. The system employs ontology with medical CIG-independent knowledge represented in OWL (so it can be processed with standard reasoners), developed in collaboration with domain experts and integrating parts of SNOMED CT and ACT terminologies. GLARE-SSCPM accepts CIGs represented in the GLARE formalism as conditional and hierarchical graphs. Similarly to GoCom, the authors adopt the mixed initiative planning paradigm where the final management plan is developed by the clinician who interacts with the system following the “focus, hypothesize and test” modality. Specifically, the system supports the clinician in focusing on relevant parts of CIGs (where adverse interactions may occur), identifying alternative management options and testing these options in “what-if” analysis. GLARE-SSCPM was tested on a case study of a patient suffering from venous thrombosis (VT) and peptic ulcer (PE).

Zamborlini et al. propose MGCDs that combines the Transition-based Medical Recommendation (TMR) knowledge representation model with first-order logic (FOL) rules. The TMR model describes CIG recommendations augmented with additional domain knowledge, such as causes and effects of actions and possible interactions between recommendations. Similarly to GoCom, recommendations are associated with goals and these recommendations may have negative, neutral or positive contributions towards the goals. FOL rules are used to identify interactions between multiple CIGs based on the knowledge encapsulated by corresponding TMR models. These models are generic and reusable, thus they do not need to be customized to specific CIGs and in this sense they are similar to integration policies introduced in CIG-IntO. The proposed approach was tested in complex case study of a patient with breast cancer and three additional multimorbidity conditions: osteoarthritis (OA), HT and congestive heart failure (CHF).

As shown in the above summary, there are multiple approaches to creating MGCDS. These approaches have diversified (but often complementary) capabilities, use different representations of the CIGs and related domain knowledge, and use different methods to develop management plans. Moreover, they were assessed using unique case studies, which makes their comparison from methodological and practical perspectives even more challenging. We believe a comprehensive comparative framework, similar to the one developed to compare CIG representations, should facilitate MGCDs comparison of functionalities, provide a common platform for presentation of various approaches, and support development of new ones.

Methods

Our methodology for identifying and confirming the features of the MGCDs consisted of three parts. First, we conducted a literature review to identify features of MGCDS used by research groups in the field. Second, we created a number of case studies that embody these features. Finally, we developed a survey whereby physicians were asked to confirm and comment on the list of identified MGCDS features.

Identification of MGCDS features
Most of the research on MGCDS is published in several health informatics journals, such as Journal of Biomedical Informatics, Journal of American Medical Informatics Association, International Journal of Medical Informatics, Methods of Information in Medicine, Journal of Medical Systems, and Artificial Intelligence in Medicine. In line with the PRISMA systematic review process\(^\text{13}\), we searched Google Scholar, PubMed and Web of Science with relevant keywords, screened the title and abstracts of the records found, assessed eligibility of the full-text, and finally reviewed the remaining publications. As a result of this review we identified 18 multimorbidity features, spread across the reviewed publications, which can be categorized as follows: (a) interactions among recommendations coming from disease-specific CIGs; (b) mitigation strategies when CIGs offer interacting recommendations; and (c) other possible features. Features from category (a) were identified based on clinical case studies presented in the literature, whereas features from categories (b) and (c) were identified based on the approaches to integrating comorbid CIGs. A complete list of features together with illustrative examples is presented in Table 1.

Case studies to demonstrate MGCDS features

All groups with prior work on MGCDS were invited to contribute relevant case studies, i.e., cases where recommendations from different clinical guidelines result in adverse interactions or introduce resource inefficiencies for a multimorbidity patient. These case studies were either based on previously published examples for demonstrating their multimorbidity decision-support methods, or represented new case studies that similarly illustrated the identification and/or mitigation of adverse interactions.

Initially, the groups were provided with a sample multimorbidity case study from Kogan et al.\(^\text{6}\) in a uniform, comprehensive format, supplemented by references to specific statements from the CPGs involved, the set of interactions to be detected and the solution – i.e., sets of treatment options that mitigate the multimorbidity interactions. We asked the groups to submit case studies in this uniform format. We selected a minimal set of submitted case studies (4) that together cover the full set of the previously identified multimorbidity features. The 4 case studies provide good coverage of the features - 9 of the 18 features are demonstrated in 2 -3 case studies. The case studies were reviewed by clinical partners for correctness. The 4 case studies are intended as a starting point to begin the comparison study of existing MGCDS; additional case studies will be added as the research progresses.

Next, we revised and expanded the selected case studies based on a rigorous process, which started with a review of the CPG repositories to identify updated versions of the guidelines and supplementary references for specific statements and actions. Collaborating medical experts were consulted for validating the clinical accuracy of the cases. As a result, we were able to establish a set of validated interactions together with the set of treatment options to be considered. Representative synthetic patient scenarios were developed with the help of medical experts and added to each of the case studies.

Validation of the MGCDS features

We started by consulting members of the groups with prior work on MGCDS to review and comment on the set of 18 MGCDS features. Secondly, we developed an online survey using the Qualtrics platform to survey physicians in order to determine the validity of the proposed MGCDS features for our framework. Physicians were recruited to complete the survey via convenience sampling. The survey preamble introduced physicians to the purpose of the study and defined notions of adverse interactions and mitigation strategies in the context of the MGCDS. The survey included 18 questions partitioned into three sections - the first section was devoted to the adverse interactions that may occur as a result of applying guidelines (7 identified features), the second section was devoted to types of mitigations to be applied when addressing such interactions (7 identified features), and the third section was devoted to other possible features of MGCDS (4 identified features). Short examples from the developed case studies were included to illustrate each feature. Physicians were asked to evaluate whether each identified feature was relevant or not for the multimorbidity problem. At the end of each section, they were provided an opportunity to add and describe any missed features. The survey was piloted with two physicians and adjustments to the phrasing of the questions was made based on their feedback. The final version of the survey can be viewed at our GitHub repository\(^\text{14}\).

Results

Case study descriptions

Twelve case studies were contributed by four of the participating groups. The minimal set of cases, which cover all identified features of MGCDS, included four cases provided by three groups and are summarized below. The full case descriptions can be accessed at our GitHub repository\(^\text{14}\).
Case 1, adapted from Kogan et al., involves three cascading morbidities. The first morbidity was managed with a drug, resulting in an adverse drug event (ADE). The ADE is regarded as another morbidity and is treated with a drug, resulting in a second ADE, which is regarded as a third morbidity. The possible mitigation strategies include either (a) adding a drug for the third morbidity; or (b) preventing one of the ADEs by replacing or stopping the drug that caused it. The various management plans may meet all clinical goals (address all current morbidities) or may compromise one of them. Specifically, Case 1 describes a patient that is on aspirin for prevention of stroke, which causes DU due to NSAID, which has been treated by stopping aspirin and adding omeprazole (a proton pump inhibitor, PPI). Aspirin was continued with the PPI to prevent DU recurrence. Now secondary osteoporosis is diagnosed, caused by the PPI.

Case 2, adapted from Michalowski et al. involves three morbidities that need to be simultaneously managed, while at the same time considering patient preferences. It describes a situation where a patient successfully treated for two concurrent conditions is diagnosed with a third one and this new diagnosis triggers the need for a revised treatment plan. The mitigation strategies include (a) making more aggressive treatment of one of the underlying conditions, and (b) managing drug contraindications and interactions. Additionally, when developing a management plan, the patient’s preferences need to be taken into account. Specifically, Case 2 describes a patient suffering from CKD and HT that are managed with ACE inhibitors, calcium channel blockers (CCB), diuretics, and low dosage aspirin (for prevention of cardiovascular disease). New diagnosis of atrial fibrillation requires the following, in line with the strategies described above: (a) replacing aspirin with an anticoagulant (warfarin) for more aggressive anticoagulant treatment, and (b) using sodium channel blockers (SCB) instead of potassium channel blockers (PCB) in anti-arrhythmic therapy (as PCB is contraindicated for the CKD patients), and abandoning beta blocker medication routinely used for rate control because of its possible interactions with ACE inhibitors or CCB. In light of patient preferences, warfarin is replaced with one of the direct anticoagulants.

Case 3, adapted from Jafarpour et al., involves two morbidities where clinical guidelines recommend adversely interacting drug treatments, both of which are nevertheless needed for treating the multimorbid conditions. The mitigation strategies include (a) increased frequency in monitoring relevant vital signs during concomitant drug treatment; and (b) adjusting drug dosage to compensate for negative evolutions of these vital signs. Further, increased frequency of monitoring must be maintained after completing one of the drug treatments until stable vital signs are observed. Specifically, Case 3 describes a patient with venous thromboembolism (VTE) and bacterial urinary tract infection (UTI) where VTE is managed by warfarin and UTI is managed by antibiotic such as trimethoprim–sulfamethoxazole (TMP/SMX). Warfarin was chosen due to availability of specific reversal agents (e.g., vitamin K); and TMP/SMX because of its low cost, effectiveness and familiarity among clinicians. During concomitant treatment, it is recommended to increase the monitoring frequency of the patient’s international normalized ratio (INR) value (e.g., daily) and adjust warfarin dosage accordingly. Upon completion of the antibiotics regimen, the increased measuring frequency should be kept in place until a stable INR is observed. At that point, regular INR monitoring should commence.

Case 4, developed especially for this study by coauthors AK, RE, MP and SWT, involves temporally managing a multimorbidity patient who needs to undergo an emergent surgical procedure. Because of the procedure, the patient has a new health risk that cannot be simultaneously addressed with other multimorbidity risks. The mitigation strategies include (a) focusing on surgery for the urgent condition; (b) suspending a long acting irreversible antagonist drug that adversely interacts with the treatment from (a); and (c) replacing it with a short acting reversible antagonist drug to minimize the time that the patient is unprotected by suspending (b). Specifically, Case 4 describes a cardiac patient with high cardiovascular risk that is on dual antiplatelet (aspirin and clopidogrel, a P2Y12 inhibitor) for 12 months following implantation of drug-eluting stent for prevention of stroke. Two months after stent implantation he is diagnosed with lung mass and needs to undergo an urgent surgical procedure that cannot be postponed past 12 months after the stent implantation—this places him at high risk for surgical bleeding due to concomitant dual antiplatelet treatment. To manage the risk, the long acting irreversible antagonist drug (clopidogrel) is suspended five days before surgery until 12-24 hours after surgery. Bridging therapy with the short acting reversible antagonist (tirofiban) is recommended. Tirofiban is started 48 hours after clopidogrel is suspended, continued until 4 hours before surgery to allow time for the drug to dissociate from platelet receptors and allow for normal aggregation and coagulation during surgery. After surgery either clopidogrel or tirofiban are resumed as soon as possible, depending on the expected degree of post operative bleeding.

Features of MGCDS
Table 1 lists and provides examples for the 18 MGCDS features. The table also points to the case studies that cover the features.

Table 1. Identified MGCDS features

<table>
<thead>
<tr>
<th>Feature</th>
<th>Short example</th>
<th>Captured by case study</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Interaction features</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A1. Drug from a CPG has an effect on a comorbid condition.</td>
<td>The cardiovascular disease CPG recommends low-dose aspirin, which may cause or worsen duodenal ulcer (DU) as a comorbid condition.</td>
<td>1,2,4</td>
</tr>
<tr>
<td>A2. Two or more drugs from different CPGs interact</td>
<td>The bacterial urinary tract infection CPG recommends antibiotics such as trimethoprim, which impacts the anticoagulant effect of warfarin that is recommended by the venous thromboembolism CPG.</td>
<td>2,3</td>
</tr>
<tr>
<td>A3. Clinical goals from different CPGs conflict</td>
<td>Coronary artery disease CPG recommend preventing thrombosis via anti-platelet therapy, which conflicts with the goal of preventing bleeding during surgery, as per perioperative antiplatelet therapy CPG.</td>
<td>4</td>
</tr>
<tr>
<td>A4. Conflicting actions (e.g., drugs, procedures) from different CPGs</td>
<td>The transient ischemic attack (TIA) CPG recommends administration of clopidogrel, while coronary artery bypass grafting CPG recommends suspending clopidogrel.</td>
<td>1</td>
</tr>
<tr>
<td>A5. Duplicate or redundant advice from different CPGs</td>
<td>Hypertension and cardiovascular disease CPGs both recommend calcium channel blockers.</td>
<td>4</td>
</tr>
<tr>
<td>A6. Temporal relationship between different CPGs</td>
<td>The acute otitis media CPG recommends taking cefpodoxime two hours after taking antacids, which are in turn recommended by the gastroesophageal reflux disease CPG.</td>
<td>4</td>
</tr>
<tr>
<td>A7. Multiple related interactions from different CPGs</td>
<td>The TIA CPG recommends aspirin, whereby the DU CPG recommends proton pump inhibitors (PPI) to mitigate the effect of aspirin on the duodenum or ulcer bleeding. PPI may cause a new comorbid condition of osteoporosis.</td>
<td>1,4</td>
</tr>
<tr>
<td><strong>Mitigation features</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A8. Adding a drug to mitigate an adverse effect</td>
<td>Add a PPI to mitigate the effect on DU caused by aspirin.</td>
<td>1</td>
</tr>
<tr>
<td>A9. Adjust drug dosage</td>
<td>A reduction of 10% of warfarin dosage to cope with concomitant treatment of antibiotics.</td>
<td>3,4</td>
</tr>
<tr>
<td>A10. Monitor the effect of a drug</td>
<td>Monitor progression of the DU during overlapping treatment with aspirin; or monitor INR frequently during concomitant treatment of warfarin and antibiotics.</td>
<td>3</td>
</tr>
<tr>
<td>A11. Replacing a drug with a safer / more effective drug for comorbidity</td>
<td>Replace aspirin with clopidogrel for a patient with DU.</td>
<td>1,2,4</td>
</tr>
<tr>
<td>A12. Discard unsafe/interacting drug</td>
<td>Suspend ACE Inhibitor when eGFR value drops by over 30% over 4 months.</td>
<td>1,2,4</td>
</tr>
<tr>
<td>A13. Delay a task to avoid a temporal overlap</td>
<td>Stop clopidogrel 5 days prior to surgery to reduce bleeding risk.</td>
<td>4</td>
</tr>
<tr>
<td>A14. Add a task to ensure a temporal overlap</td>
<td>When stopping clopidogrel prior to surgery, start bridging therapy with tirofiban 24h later until 4h before surgery, and resume 2h after surgery.</td>
<td>4</td>
</tr>
<tr>
<td><strong>Other features</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A15. Patient preferences and/or patient burden</td>
<td>Choosing one drug over another due to lower price; or choosing any of direct oral anticoagulants over warfarin to avoid checking INR on regular basis.</td>
<td>1,2,3,4</td>
</tr>
</tbody>
</table>
A16. Optimization of clinical resources
Grouping tests recommended by different CPG on the same day, or avoiding multiple imaging scans, recommended by different CPG, where results can be re-used for diagnosis of both comorbid illnesses.

A17. Explanation of the mitigation strategy(ies)
Including an explanation for a recommended mitigation (e.g., all patient conditions are treated, the largest number of conditions are treated, or the condition that is at the focus of the medical investigation is treated).

A18. Alternative mitigation strategies for a single interaction
For a patient taking aspirin for secondary prevention of TIA, who developed DU due to aspirin, one strategy may be to add a PPI to protect the duodenum, and a second strategy may be to replace aspirin with clopidogrel.

Validation of MGCDS features
Members of all groups with prior work on MGCDS were asked to review the identified MGCDS features and to suggest any missing ones. The 11 group members who responded were positive about the set of 18 proposed features and did not suggest any new ones. After this initial validation, we developed and validated an online survey and recruited 15 physicians of different specialties and different levels of experience for assessing and commenting on the features. The survey was completed by all invited physicians. The results are presented in Table 2.

Overall, the results of the survey confirmed the relevance of the identified MGCDS features. There were only a few instances where physicians did not endorse the features unanimously and these are outlined here. Regarding features associated with interactions among recommendations coming from disease-specific CPGs, identification of duplicate or redundant advice from different CPGs (A5) was found to be relevant by 9 out of 15 physicians while identifying temporal relationship between different CPGs (A6) and identifying conflicting actions from different CPGs (A4) were found to be relevant by 13 and 14 out of 15 physicians, respectively. Regarding features associated with the mitigation strategies when CPGs offer interacting recommendations, the mitigation strategies of monitoring the effect of a drug (A10) and replacing a drug with a safer/non-interacting drug/more effective drug for comorbidity (A11) were found to be relevant by 13 and 14 out of 15 physicians, respectively. The least agreement among the physicians was observed for the other possible features’ category. Here, only identification of alternative mitigation strategies for a single interaction received unanimous support. Fewer physicians were convinced that explanation of the mitigation strategy(ies) (A17) with 11 positive responses out of 15, optimization of clinical resources (A16) with 12 positive responses out of 15, and inclusion of patient preferences and/or patient burden (A15) with 14 positive responses out of 15 are relevant.

Table 2. Physician responses to survey

<table>
<thead>
<tr>
<th>Features of the multimorbidity CPG problem</th>
<th>#Physicians who found the features relevant (out of 15)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interactions among CPGs’ advice</td>
<td></td>
</tr>
<tr>
<td>A1. Drug from a CPG has an effect on a comorbid condition</td>
<td>15</td>
</tr>
<tr>
<td>A2. Two or more drugs from different CPGs may interact</td>
<td>15</td>
</tr>
<tr>
<td>A3. Clinical goals from different CPGs may conflict</td>
<td>15</td>
</tr>
<tr>
<td>A4. Conflicting actions (e.g., drugs, procedures) from different CPGs</td>
<td>14</td>
</tr>
<tr>
<td>A5. Duplicate or redundant advice from different CPGs</td>
<td>9</td>
</tr>
<tr>
<td>A6. Temporal relationship between different CPGs</td>
<td>13</td>
</tr>
<tr>
<td>A7. Multiple related interactions from different CPGs</td>
<td>15</td>
</tr>
<tr>
<td>Mitigation strategies when CPGs offer interacting advice</td>
<td></td>
</tr>
<tr>
<td>A8. Adding a drug to mitigate an adverse effect</td>
<td>15</td>
</tr>
</tbody>
</table>
Physicians did not indicate that any features were missing from the set provided for evaluation. Physicians made few comments mostly related to prioritizing goals from CPGs: one suggestion was to ignore actions that are associated with less important goals and prioritizing goals based on clinical needs. Another physician suggested that goals should be prioritized based on what treatment a patient can or is willing to follow. Finally, one physician commented that the most difficult aspect of MGCDS is an assessment of risks and benefits when guidelines are in conflict.

**Discussion and Future Work**

Multimorbidity is complex clinically but also challenging for effective decision support. This challenge is manifested by a relatively large number of published MGCDS, with none of them covering all possible features associated with supporting the management of multimorbidity patients. Therefore, it is important that there is a unified framework that, on the one hand, allows for comparing functionalities of existing MGCDS, and, on the other hand, can help guide development of new ones by highlighting gaps in the state-of-the-art. The purpose of the research described in the paper was to create such a framework. We have identified a set of MGCDS features, developed 4 case studies to cover those features and conducted a survey with physicians to confirm the features.

The survey results largely confirmed the feature set. There were three features where relevance was somehow questionable for the physicians. Five respondents did not consider *duplicate or redundant advice of different CPGs* (A5) as a relevant feature. The most plausible explanation is that experienced clinicians find such advice to be rather straightforward. However, considering that an MGCDS might be used by physicians of different levels of experience, having such a feature may be useful. Similarly, three physicians considered *explanation of mitigation strategy(ies)* (A17) to be less relevant. Such thinking seems to be related to the ongoing discussion in the medical informatics community about the “black box algorithm effect”, with some arguing for system explainability while others focus on the quality of performance of a black box algorithm. It is our assertion that the assessment of this feature reflects this debate. Finally, two physicians asserted that *optimization of clinical resources* (A16) is not relevant. In our context, optimization of the resources implies avoiding unnecessary tests or grouping these tests together so they can be conducted during one visit. While this is probably one of the most relevant features from a patients’ perspective, physicians consider such optimization to be beyond scope of their practice and being under control of laboratory and imaging services.

The strengths of our method to develop this comparative framework includes the thorough review and analysis of the existing MGCDS literature, the participation of original developers of various guideline-based multimorbidity methods, the rigorous vetting of the cases by physicians, the confirmatory survey by physicians not involved in the development process, and the upcoming comparison study that will use, refine, and extend the framework. At the same time, we recognise certain limitations. Firstly, our sample size of physicians is small (15 physicians), however, we believe it achieves the goal of verifying the validity of features at this point in the work, and we intend to validate the MGCDS framework with a larger group of physicians in the upcoming phases of our work. Second, given that we
derived the set of interaction, mitigation, and other features from a review of the existing MGCDS literature, it is possible that additional features may be discovered as researchers work on new domains and new combinations of morbidities. Thus, this framework will necessarily be an evolving one and will merit future reviews. Finally, another limitation of our method is that we did not mine the clinical literature for potential sources of new features. While many of us were inspired by the landmark 2005 Boyd et al. paper, it is beyond our expertise and scope to review the clinical literature for interesting and novel multimorbid interactions and mitigations. Nevertheless, given the rigorous review of the case studies and the affirmation of the features by physicians—whereby no additional features were suggested, with one caveat (see below)—we are confident about the robustness of the case studies and the multimorbidity features that we have identified. It is our hope that this comparative framework and upcoming study will be of interest not only to informaticians, but to clinicians as well. With a consolidation of existing understanding of multimorbidity interactions and mitigations, we will be in the position to have further dialog between informaticians and clinicians.

Another aspect of our framework is that it is more methodologically-oriented than implementation-oriented—meaning that it is focused on high-level features and mitigation strategies rather than concrete implementation and deployment methods. What the groups contributing to this framework share is that they have developed MGCDSs that in totality cover the identified features. Originally some of us hoped that the framework and the survey could shed new light on the requirements of implementing MGCDS as well. However, the implementation of MGCDS as actual systems for deployment depends on myriad factors (e.g., the target audience and the workflow settings) beyond what the framework can accommodate. The downgrading of explainability and optimization of clinical resources by some physicians may also be a reflection of this issue.

The clinicians’ comments about the need to prioritize conflicting goals or weighing risks and benefits of different actions suggest that a new “other feature” may possibly be relevant, which focuses on explicit support for decision making among conflicting goals and actions. Some approaches in the literature already provide such decision support. For example, MitPlan tries to optimize an objective function (e.g., overall cost) that is selected by the clinician. GoCom makes the choices among different alternatives explicit, some of which may not satisfy a given guideline-suggested goal, but does not support weighing the priorities, costs and benefits, and trade-offs among the alternatives.

We plan to complete (3) a reporting standard for MGCDS solutions and (4) criteria for evaluating MGCDS solutions of our framework as part of the upcoming comparison study. In this study, we will use the complete comparative framework to evaluate the existing MGCDSs with direct involvement from groups that designed the systems. A quantitative evaluation will assess functionalities of MGCDS and in a qualitative evaluation we will interview physicians about the MGCDS solutions. We will also use a larger set of case studies, including real world case studies, investigate further conflicts and how to mitigate these conflicts, and explore the identification of further possible MGCDS features with physicians. A companion paper describing comparison study will present the development of part 3 and part 4 of the framework as well as the study results themselves. A reporting template will necessarily incorporate some evaluation criteria of the MGCDS systems themselves (e.g., the use of standard terminologies or knowledge sources). Several groups are already piloting a reporting template proposal, using not only their own cases but also external cases as exemplars. We expect to present the piloted reporting templates and a few reporting exemplars to the groups participating in the comparison study and iteratively refine them to the groups’ satisfaction. Once the comparison study starts, we expect each group to implement guideline fragments sufficient to execute the common cases to the extent possible and then to report on their results.

To summarize, the results described in this paper represent first steps towards creating a validated, comprehensive framework for comparing functionalities of MGCDS. Having such a framework should help with identifying gaps in MGCDS research and subsequently help with moving this research area forward. In order to facilitate this progress, we plan to prospectively evaluate our proposed framework by inviting different research groups working on MGCDS to use the framework and its accompanying clinical use cases. This should help with identifying gaps in MGCDS research as well as provide guidance for future research directions.

References

Predictive and Causal Analysis of No-Shows for Medical Exams During COVID-19: A Case Study of Breast Imaging in a Nationwide Israeli Health Organization

Michal Ozery-Flato, PhD¹, Ora Pinchasov, BSc², Miel Dabush-Kasa, BSc², Efrat Hexter, BSc¹, Gabriel Chodick, PhD³,⁴, Michal Guindy, MD, MPA²,⁵, Michal Rosen-Zvi, PhD¹,⁶

¹IBM Research-Haifa, Haifa, Israel; ²Assuta Medical Centers, Tel Aviv, Israel; ³Maccabi Healthcare Services, Tel Aviv, Israel; ⁴Tel Aviv University, Tel Aviv, Israel; ⁵Ben Gurion University of the Negev, Beer Sheva, Israel; ⁶The Hebrew University, Jerusalem, Israel

Abstract

“No-shows”, defined as missed appointments or late cancellations, is a central problem in healthcare systems. It has appeared to intensify during the COVID-19 pandemic and the nonpharmaceutical interventions, such as closures, taken to slow its spread. No-shows interfere with patients’ continuous care, lead to inefficient utilization of medical resources, and increase healthcare costs. We present a comprehensive analysis of no-shows for breast imaging appointments made during 2020 in a large medical network in Israel. We applied advanced machine learning methods to provide insights into novel and known predictors. Additionally, we employed causal inference methodology to infer the effect of closures on no-shows, after accounting for confounding biases, and demonstrate the superiority of adversarial balancing over inverse probability weighting in correcting these biases. Our results imply that a patient’s perceived risk of cancer and the COVID-19 time-based factors are major predictors. Further, we reveal that closures impact patients over 60, but not patients undergoing advanced diagnostic examinations.

Introduction

The outbreak of the coronavirus disease 2019 (COVID-19) pandemic in early 2020 led to a unique unprecedented societal shift. Governments implemented a wide variety of nonpharmaceutical interventions (NPIs) aimed at controlling the disease spread and enforcing social isolation. Confinement and school closure as well as entertainment and cultural sector closure were the predominant NPIs taken by governments in 2020. Accordingly, many radiology departments decreased the number of elective imaging examinations to minimize the spread of infection and free up much needed medical resources and staff. One specific example is the reduction of breast imaging to 28% of baseline imaging volumes reported by Stanford Health Care during the first months of the pandemic outbreak. Another is the 85% drop in mammography examinations over five days during March 2020 as reported by a large, metropolitan hospital system consisting of six outpatient practices across three New York City boroughs. Many radiology centers in the United States started to gradually reopen these examinations in the summer of 2020. In parallel to the variability in the outpatient radiology practices regarding examination availability, there was also a change in the behavior of patients coming to the exams, showing a distinct rise in no-show rate that seemed to be correlated with the height of the pandemic.

No-shows to medical appointments significantly impact revenue, cost, and the efficient use of resources. Understanding this issue can help us to better manage resources, target interventions to prevent the no-show, and reduce some of the associated revenue loss and cost. Predicting patient no-shows and identifying the major individual attributes associated with no-show have been the key approach to understanding this phenomenon. Patient cohorts with retrospective data derived from electronic health records (EHRs) have been leveraged to train prediction models, with logistic regression being the most widely-used model. Specifically, among the different modalities used in radiology examinations, mammography (MG) and CT were identified as having the highest no-show rates. Among the different predictors, age and scheduling lead time were often identified as highly associated with not showing up.

In recent years, scalable methods were developed to better understand different phenomena by applying causal inference to patient cohorts. Causal inference can help quantify the magnitude of the effect of specific interventions on no-shows. In the context of the pandemic and the NPIs imposed by governments, it is important to assess the effect of NPIs on the no-shows. Following the common design of causal inference studies, we analyze two groups in the data: appointments that took place when NPIs were not imposed and appointments that took place during the time of imposed NPIs. Inverse Probability Weighting (IPW) is a method frequently used to correct for unequal confounders distribution across the groups. It is aimed at eliminating biases in the data and enabling an inference of the average
effect at the population level\textsuperscript{9}. However, IPW is not guaranteed to successfully remove biases. A recently developed method called Adversarial Balancing (AdvBal)\textsuperscript{10}, based on generative adversarial networks (GANs), is better posed to address covariates biases in high-dimensional data and generate weights that are closer to uniform. We tested and compared IPW and AdvBal to estimate the average effect of the NPIs on the no-show rate.

In this study we analyze the data of patients coming from the largest private network of hospitals in Israel, which performs 30\% of the screening mammography exams in the country. Excluding several days, which were dismissed from our no-shows analysis, this network did not apply any policy to decrease the number of imaging exams during the COVID-19 pandemic, but rather adjusted service to changes in the demand. We present a comprehensive study of no-shows during COVID-19, taking into account time-based coronavirus factors such as \% confirmed cases, reproduction rate, NPIs, and changes in population mobility, which are derived from multiple data resources. See Figure 1 for the trend of selected time-based statistics over the year 2020. The correlations between changes in the growth of COVID-19 morbidity, population mobility, and the ratio of show-to-no-show are easily noticed. To the best of our knowledge, this is the first study to combine pre-pandemic known predictors of no-show with pandemic associated predictors. Our study is also unique in the causal inference analysis that we perform to assess the average effect of NPIs on no-show in the entire study population, and in subgroups of interest.

**Methods**

Our study analyzes no-show appointments over the course of 2020. We refer to the time period from January 3 till February 6, 2020 as the *Pre-COVID-19 period*, and to March 15 till December 31, 2020 as the *COVID-19 period*. The Pre-COVID-19 period was defined in a manner similar to the baseline period used in Google Community Mobility Reports\textsuperscript{11}. We focused on the COVID-19 period and used the Pre-COVID-19 period as a reference for its comparison. The inspection of NPI data revealed large correlations in these data, as expected. Therefore, we decided to include in our analysis a single NPI, school closure, which is more easily defined, and serves as an indication for other NPIs. Similar to previous studies\textsuperscript{12,13}, we consider late cancellations as no-shows since they do not allow the scheduling
system to set an appointment for a new patient. We randomly partitioned the patients in this study into train (56%), validation (14%), and test (30%). We trained our prediction models on the train dataset and evaluated them on the validation dataset during the development process. The train and validation sets were also used for the development of the causal analysis. The test dataset was held-out until the model, analysis, and all hypotheses were finalized. Unless specified otherwise, the results presented in this study are from the held-out data.

Data

This retrospective study was approved by the institutional review board of Assuta Medical Center (0033-20-ASMC), who waived the requirement for patient consent. The appointments dataset was extracted from the EHR datamart of Assuta Medical Centers and included all imaging appointments from January 1, 2020 to December 31, 2020. Our initial cohort included all patients with mammography (MG) and/or breast ultrasound (BUS) appointments during 2020. In addition to the scheduled date, the information on each appointment included the creation and cancellation dates, with the latter filled only for cancelled appointments. Rescheduling an appointment involves cancelling the old appointment and creating a new one, and therefore all dates in an appointment record are assumed non-modifiable. The information on the scheduled exam included: the imaging modality (MG, BUS), the procedure code, a pre-exam urgency indicator, and the site of the exam. The information on the patient included: a (de-identified) patient-ID, gender, age, and town of residence. We received additional data on each patient selected for our study, which included the results of the patient’s last MG, BUS, breast MRI, and breast biopsy exams, if present.

We obtained socio-economic measures and population sizes for localities in Israel from the Central Bureau of Statistics (CBS) reports. We downloaded Israel Transverse Mercator (ITM) coordinates for localities in Israel from the public Israeli Government Databases and used these data to compute pairwise distances between the patients’ towns and exam sites. Time-based statistics of coronavirus morbidity in Israel, and for specific towns, were downloaded from the COVID-19 Data Repository of the Israeli Ministry of Health. We obtained data on NPIs in Israel from the IBM Worldwide Non-pharmaceutical Interventions Tracker for COVID-19 (WNTRAC). The data included information on confinements, school closures, public services closures, work restrictions, and more. Finally, we acquired community mobility data in Israel from Google Community Mobility Reports. These reports include the per-day change in movement across six different categories of places, with respect to the median value of the same day of the week during the five-week period January 3 to February 6, 2020.

No-Show Outcome

We defined a no-show outcome for a pair of (patient, date), if the patient had an appointment for an imaging exam on that date. There were many patients with multiple appointments on the same date, e.g., MG and BUS for breast cancer screening were often scheduled for the same date. We classified a pair of (patient, date) as show if the patient attended at least one appointment at that date, and no-show if the patient attended none. Our assumption was that partial attendance to appointments is likely not the patient’s choice. We inferred that a site had no service for a certain modality on a specific date, if no corresponding exams were performed on that date. We then refined the no-show definition to exclude (patient, date) pairs for which there was no service for all patient appointments on that date.

Cohort Selection and Index Date

Our initial cohort included all (patient, date) pairs during the year 2020, where the patient had an ambulatory appointment for an MG or BUS exam on that date. We included only patients aged 18 years or older. We set the index date of each (patient, date) sample to be one-week (7 days) prior to that date, and discarded (patient, date) samples for which all appointments were created at or after the index date. We referred to these discarded samples as “same week schedules” and analyzed them in a separate study, as these were shown to have distinct characteristics. We also excluded (patient, date) samples for which all appointments were cancelled prior to the index date. We refer to such samples as “early cancellations” and assumed the health system can accommodate for these. To summarize, we selected (patient, date) samples with at least one planned appointment by the index date, by excluding samples that had all their appointments cancelled at that date, or not yet created. The process of cohort extraction is depicted in Figure 2.

Features

We used the data from the time period prior to the index date to extract features for our no-show prediction model. The same set of features was used in the analysis of closure effects. We extracted a total of 140 features, using a previously described feature extraction tool. Below we present these features:

Patient demographics: age, gender, town socio-economic index-value, town distance from site.
Figure 2. Cohorts’ inclusion/exclusion criteria

Index exams: waiting time (time between the date of the appointment creation and the scheduled date, aka “lead time”), imaging modality, total number of modalities, MG/BUS procedure indicator, total number of MG/BUS procedures, pre-exam urgency indicator, site indicator.

Imaging history (3 months): total number of show dates (all imaging modalities), total number of show dates for the same modality, total number of no-show dates (all modalities), total number of no-show dates for the same modality.

Cancer diagnostic phase: we mapped each MG/BUS procedure to a cancer diagnostic phase. Basic and extended breast cancer screening procedures were mapped to phases 1 and 2 respectively; diagnostic exams of patients recalled after screening exams were mapped to phase 3; phases 4 and 5 pertain to biopsy-related procedures, indicating the existence of findings suspected as potentially malignant. The feature “is screening” indicates basic screening (phase 1).

Previous MG and BUS exams: has a previous exam indicator, time to last exam, last procedure indicator, last cancer diagnostic phase, last procedure is screening, last BI-RADS score (i.e. the result of the exam), last breast density measure (1-4).

Previous breast MRI and biopsy exams: has a previous MRI indicator, time to last MRI, MRI procedure indicator.

COVID-19 related measures: # new confirmed cases, % confirmed cases, change in population mobility for each of the six place categories, school closure indicator, reproduction rate. All these features refer to measures at the country-level. We also extracted town-specific time-based features corresponding to the patient’s town: total number of new confirmed cases in town (per 10K), % confirmed cases in town. All time-based features were computed after the underlying time-series data were smoothed with a moving average of seven days. The COVID-19 reproduction rate was inferred with the EpiEstim\textsuperscript{19} application.

Predictive Analysis of No-Shows

We trained our no-show prediction model with the gradient boosted trees algorithm, using the XGBoost\textsuperscript{20} package. Hyperparameters were tuned with 5-fold cross validation on the train dataset, via a randomized grid search. We were unable to map the town of residence to the list of towns published by the Israel CBS for 7% of the samples, with a higher rate for no-show samples. Inspecting the unmatched town values revealed that most of these corresponded to either “unknown” or “other” values. We preprocessed the train data and imputed missing values in town-based features, to prevent XGBoost from learning predictive patterns of missing values in these features. The imputation was done with the stochastic “multivariate imputation by chained equations” (MICE) method, with ten iterations, using the “Statsmodels”\textsuperscript{21} package. We relied on the inherent XGBoost imputation mechanism to handle missing values in the remaining features during train, and in all the features during inference. We used TreeExplainer from the Shapley Additive exPlanations (SHAP)\textsuperscript{22} package to estimate the contribution of each feature to our XGBoost
model, based on the classic game-theoretic Shapley values. We obtained an agglomerative hierarchical clustering of time-series data using average linkage, and the Euclidean distance between the pairwise Pearson correlations as the distance metric.

**Causal Analysis of Closures**

We divided our (patient, date) samples into closure and no-closure groups based on whether there was a school closure on the date. We leveraged the average treatment effect (ATE) to examine the effect of school closure on no-show rate:

\[
ATE = \mathbb{E}(Y_{\text{closure}}) - \mathbb{E}(Y_{\text{no closure}})
\]

(1)

where \(\mathbb{E}(Y^1)\) and \(\mathbb{E}(Y^0)\) are the potential no-show rates for the entire cohort if all samples were during closure and no-closure, respectively. In other words, the ATE is the average effect, at the population level, of moving the entire study population from no-closure to closure. Estimating the ATE from the observed data requires accounting for confounding biases between closure and no-closure groups. For each feature, we measured the bias between the two, possibly weighted, closure and no-closure groups by computing the absolute standardized difference:

\[
d = \frac{|\bar{x}_{\text{closure}} - \bar{x}_{\text{no closure}}|}{\sqrt{(s^2_{\text{closure}} + s^2_{\text{no closure}})/2}}
\]

(2)

where \(\bar{x}_{\text{closure}}\) and \(\bar{x}_{\text{no closure}}\) are the feature means in the closure and no-closure groups, respectively, and \(s^2_{\text{closure}}, s^2_{\text{no closure}}\) are the corresponding sample variances. A standardized difference that is less than 0.1 is commonly regarded as negligible bias between compared groups. We visually compared the distributions of top-biased features within the closure and no-closure groups. We plotted the smoothed densities of the compared distributions using Gaussian kernel density estimation with a bandwidth of 1.

**Inverse probability weighting (IPW)**. IPW is a longstanding popular method that weights samples based on the inverse of their propensity to be in their assigned group. The propensity is estimated by fitting a model that predicts the assigned group. We used a logistic regression model, which is commonly used for IPW.

**Adversarial balancing (AdvBal)**. The AdvBal method borrows principles from GANs, to generate sample weights for a source data, such that the resulting weighted data becomes similar to a given target data. In our case, it was aimed at generating weights for the closure (resp., no closure) group, to make it indistinguishable from the entire cohort data. Similar to GANs, it operates by alternately training a classifier to distinguish between the source and target data, and a weights generator that uses exponentiated gradient descent to maximize classification error. We applied AdvBal with a logistic regression classifier. We set the number of iterations in AdvBal to ten.

Prior to applying logistic regression modeling to the data, in both IPW and AdvBal, we imputed missing values in all the features using the MICE method. We used the implementation of IPW and AdvBal from the CausalLib package.

**Evaluation Measures and Statistical Tests**

We measured the accuracy of our prediction models using the area under the receiver-operator curve (AUC). We assessed the statistical association between two binary variables with a chi-squared test, and between a binary variable and a non-binary variable with an unpaired one-sided t-test. These univariate statistical tests were performed after omitting missing values in the tested features. We compared two AUCs that were computed on the same data using the DeLong test. We estimated the standard error of computed effects with 100 bootstrapping iterations, and computed P-values by the normal distribution. We used the validation set for testing various hypotheses, from which we selected 25 that seemed most attainable. We applied the Benjamini-Hochberg procedure to account for multiple testing, and all the 25 P-values reported in this study were significant at false discovery rate lower than 0.002.

**Results**

Table 1 presents selected statistics for the train, validation, and held-out datasets for each of the COVID-19 and Pre-COVID-19 periods. As shown, the train, validation and held-out datasets exhibit similar statistics for both periods. A comparison of the COVID-19 and Pre-COVID-19 periods revealed that the no-show rate significantly increased during COVID-19 (35% vs. 31%, \(P\)-value = 0.0002, chi-squared test), due to the increase in late cancellations (11% vs. 7%). Note that late cancellations also include late reschedules, that is, reschedules during the week before the appointment. Another prominent difference between the two time periods is the shorter average waiting time during COVID-19 (41 vs 44 days, \(P\)-value = \(2 \times 10^{-16}\)), which can be explained by the 7% reduction in the average
number of appointments per week. Both “%confirmed cases” and “school closure” showed a significant association with no-shows during COVID-19.

Table 1. Cohort statistics. For binary variables, we present the proportion (%) and mean ± standard deviation for non-binary features. The association with the no-show outcome (show and no-show statistics, P-value) is presented within squared brackets.

<table>
<thead>
<tr>
<th>Time</th>
<th>Feature</th>
<th>Train</th>
<th>Validation</th>
<th>Held-out</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N; N / no. patients</td>
<td>10,427; 1.04</td>
<td>2,524; 1.03</td>
<td>5,517; 1.04</td>
</tr>
<tr>
<td>Pre-COVID-19</td>
<td>No-show: all; late cancellation</td>
<td>31 %; 7%</td>
<td>30 %; 8%</td>
<td>32 %; 7%</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>99 %</td>
<td>99 %</td>
<td>99 %</td>
</tr>
<tr>
<td></td>
<td>Age</td>
<td>55±13</td>
<td>55±12</td>
<td>55±13</td>
</tr>
<tr>
<td></td>
<td>Breast cancer screening</td>
<td>62 %</td>
<td>62 %</td>
<td>63 %</td>
</tr>
<tr>
<td></td>
<td>Cancer diagnostic phase</td>
<td>2.2±0.6</td>
<td>2.2±0.6</td>
<td>2.2±0.6</td>
</tr>
<tr>
<td></td>
<td>Waiting time (days)</td>
<td>44±28</td>
<td>45±28</td>
<td>44±28</td>
</tr>
<tr>
<td></td>
<td>Has previous mammography</td>
<td>70 %</td>
<td>70 %</td>
<td>71 %</td>
</tr>
<tr>
<td></td>
<td>Time to last mammography (years)</td>
<td>1.7±0.8</td>
<td>1.7±0.8</td>
<td>1.7±0.8</td>
</tr>
<tr>
<td>COVID-19</td>
<td>N; N / no. patients</td>
<td>80,362; 1.27</td>
<td>20,181; 1.27</td>
<td>43,487; 1.27</td>
</tr>
<tr>
<td></td>
<td>No-show: all; late cancellation</td>
<td>35 %; 11%</td>
<td>34 %; 11%</td>
<td>35 %; 11%</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>99 %</td>
<td>99 %</td>
<td>99 %</td>
</tr>
<tr>
<td></td>
<td>Age</td>
<td>55±13</td>
<td>55±13</td>
<td>55±13</td>
</tr>
<tr>
<td></td>
<td>Breast cancer screening</td>
<td>61 %</td>
<td>61 %</td>
<td>61 %</td>
</tr>
<tr>
<td></td>
<td>Cancer diagnostic phase</td>
<td>2.2±0.6</td>
<td>2.2±0.6</td>
<td>2.2±0.6</td>
</tr>
<tr>
<td></td>
<td>Waiting time (days)</td>
<td>41±30</td>
<td>42±31</td>
<td>41±30</td>
</tr>
<tr>
<td></td>
<td>Has previous mammography</td>
<td>72 %</td>
<td>73 %</td>
<td>72 %</td>
</tr>
<tr>
<td></td>
<td>Time to last mammography (years)</td>
<td>1.7±0.9</td>
<td>1.7±0.9</td>
<td>1.7±0.9</td>
</tr>
<tr>
<td></td>
<td>Confirmed cases per day</td>
<td>0.05±0.03</td>
<td>0.05±0.03</td>
<td>0.05±0.03</td>
</tr>
<tr>
<td></td>
<td>School closure</td>
<td>25 %</td>
<td>25 %</td>
<td>26 %</td>
</tr>
</tbody>
</table>

No-Show Predictors During COVID-19

We examined the top-20 contributing features for our XGBoost model that was trained on COVID-19 data with the entire set of 124 extracted features (Figure 3a). The strongest predictor was “cancer diagnostics phase”, an ordinal feature with 5 categories, ranging from basic screening to biopsy related procedures. Additional related features that appear in this list of top-20 were: “last MG cancer diagnostic phase”, and “screening mammography”. The features “last MG BI-RADS score” and “last BUS BI-RADS score” are also related to “cancer diagnostics phase” since BI-RADS scores determine whether patients are recalled for the next diagnostic phase. An inquiry on “procedure BUS 760904”, which was classified as cancer diagnostic phase 2, revealed that it was commonly used when the patient co-scheduled a basic MG screening exam. Thus 6 of the top-20 important features pertain to the patient’s knowledge, or perception, of the a-priori chance of cancer diagnosis. The inspection of “time to last MG”, which was ranked second among the top-20 features, revealed that patients who had mammography exams approximately a year before had the highest tendency to show up on-time to their appointments. However, this tendency decreased as the time from the previous mammography increased.

Six out of the top-20 important features were COVID-19 related factors. The exclusion of these features from our model significantly reduced the AUC from 0.744 to 0.710 (P-value = 2 × 10⁻⁰⁹⁰, DeLong test). Our results imply that when there is a growth in COVID-19 morbidity, patients may be more reluctant to attend MG and BUS exams, as reflected by the positive trend of “% confirmed cases” and “reproduction rate” in Figure 3a. The time-series of COVID-19 measurements largely correlated with each other, and a clustering analysis of these data revealed two clusters that are negatively correlated with each other (see Figure 3b). The correlation of school closure with mobility changes, which were included in list of top-20 important features, may explain the exclusion of the former from that
Estimated School Closure Effect on No-Shows

We analyzed the average effect of school closures on no-show rate, after correcting for observed confounding biases between the two time periods. As mentioned above, school closure was selected as a representative of a larger set of highly correlated NPIs and its estimated effect corresponds to the entire set of co-occurring NPIs. We estimated the average effect of school closure in our entire study population, as well as for subgroups defined by (i) age and (ii) cancer diagnostic phase. Our set of potential confounders included all extracted features, excluding school closure itself, and the reproduction rate and mobility features, which were highly correlated with the school closure indicator (See Figure 3b). We tested and compared two causal inference methods for correcting observed biases: IPW and AdvBal. We used a logistic regression model in both methods.

Comparison of IPW and AdvBal. In all our analyses, AdvBal substantially diminished the standardized difference of all the confounders (d ≤0.02). Reducing the number of the iterations in AdvBal to 5 instead of 10 also successfully reduced biases below the required threshold of 0.1 (d ≤0.08). On the other hand, IPW failed to reduce major
imbalances, as shown in Figure 4; it even introduced much larger biases into confounders that had negligible biases in the original (unweighted) data.

Estimated effects. Table 2 presents the uncorrected and corrected effects of school closure on no-show rates, in the entire study population (“All”) and for subgroups by age and cancer diagnostic phase. As shown, the estimated effect practically vanished in the subgroup of patients with advanced cancer diagnostic exams (phases 3-5), and the remaining estimated effects also largely dropped, with a decrease of 60% to 80% of the original (uncorrected) effects. As an example, the estimated average effect in the entire study population dropped from 12% to 3% after accounting for the observed confounding biases. Among tested subgroups, patients at age 60+ were most affected by closures, with the estimated potential no-show rate lifting from 31% to 37%.

Discussion

This paper presents a comprehensive analysis of no-shows, defined as missed appointments or late cancellations, during the COVID-19 pandemic, systematically verifying previously made hypotheses. Measures of COVID-19 growth, such as “%confirmed cases”, “new confirmed cases”, and the reproduction rate, are often published in the Israeli media. Our results suggest these representations of the perceived severity of the pandemic play a major role in a patient’s decision of whether to attend a medical appointment. During 2020, there were two periods of school closures in Israel, in which additional non-pharmaceutical interventions were applied, such as confinements and bans of mass gatherings. These closures covered 25% of the appointments in our data during the COVID-19 period (see Table 1). We demonstrated that the observed large effect of school closures on no-show rate dramatically decreased by 60% to 80% when accounting for biases in COVID-19 growth measures. Moreover, we show that closure effect was magnified for patients over 60 years, presumably due to their elevated risk for coronavirus complications. Conversely, no effect was observed for patients in advanced cancer diagnostic phases. The adjustment for the observed biases was enabled by the adversarial balancing method, after the application of the popular IPW method failed to balance these biases. Adversarial balancing harnessed the power of GANs to generate balancing weights that made the multivariate distributions of the confounders in the closure and non-closure groups less diverged. In this study, we

![Figure 4. Distribution of top-biased confounders during closure and no-closure periods.](image) Plots depict smoothed estimated densities for the 5 biased features in the original data (left column, d>0.1), and after reweighting the data with IPW (middle column, d>0.1) and AdvBal (right column, d<0.007). The standardized difference d (see Methods section), which measures the bias between the closure and no-closure groups, appears at the upper right corner of each subplot. Green and red colors indicate negligible (≤0.1) and non-negligible (>0.1) values for d, respectively.
applied adversarial balancing with a simple logistic regression classifier, but stronger deep learning classifiers can be used to increase its power.

**Table 2. Uncorrected and corrected average closure effects on no-show rate.** Estimated effects are presented with the estimated %no-shows for no-closure and closure groups. Negligible (d<0.1) and non-negligible(d>0.1) standardized difference values are colored in green and red, respectively.

<table>
<thead>
<tr>
<th>Group</th>
<th>N</th>
<th>% of cohort</th>
<th>%no-show</th>
<th>% in closures</th>
<th>Uncorrected Estimated effect</th>
<th>Max. bias (d)</th>
<th>Corrected (AdvBal) Estimated effect</th>
<th>Max. bias (d)</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>43,487</td>
<td>100%</td>
<td>35%</td>
<td>26%</td>
<td>12%: 32%↑43%</td>
<td>0.65</td>
<td>3%: 33%↑36%</td>
<td>0.01</td>
</tr>
<tr>
<td>Cancer diagnostic phase</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1-2</td>
<td>35,263</td>
<td>81%</td>
<td>38%</td>
<td>26%</td>
<td>13%: 34%↑47%</td>
<td>0.65</td>
<td>4%: 36%↑40%</td>
<td>0.02</td>
</tr>
<tr>
<td>3-5</td>
<td>8,224</td>
<td>19%</td>
<td>21%</td>
<td>24%</td>
<td>6%: 20%↑26%</td>
<td>0.67</td>
<td>0%: 21%→21%</td>
<td>0.02</td>
</tr>
<tr>
<td>Age (years)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;60</td>
<td>27,927</td>
<td>64%</td>
<td>35%</td>
<td>26%</td>
<td>10%: 32%↑42%</td>
<td>0.65</td>
<td>2%: 34%↑36%</td>
<td>0.02</td>
</tr>
<tr>
<td>&gt;=60</td>
<td>15,560</td>
<td>36%</td>
<td>34%</td>
<td>25%</td>
<td>15%: 30%↑46%</td>
<td>0.66</td>
<td>6%: 31%↑37%</td>
<td>0.01</td>
</tr>
</tbody>
</table>

Our results are in agreement with existing literature on no-shows5,7: patient’s historical behavior, wait time (aka “lead time”), distance from the exam site, and age were found to be predictive for no-shows. Age was also found to be predictive for no-shows but with mixed effects due to several reasons. First, breast screening in Israel is recommended at ages 50 to 74, and hence women under 40 are more likely to attend MG/BUS exams due to some indication, or suspicion, for elevated risk of breast cancer. In addition, in the screening population (diagnostic phase 1), the cancer risk is known to increase with age; however, patients at age 60+ in our data were at risk for coronavirus complications, which may have caused these patients to refrain from attending medical exams. We also demonstrated that including per-site indicators in the model can significantly improve the model accuracy, as it allows the model to identify predictive patterns that are site-specific. Finally, an important finding of our study is the inclusion of the cancer diagnostic phase as a predictor in our model. This predictor, which may reflect cancer risk as perceived by the patient, was ranked as most influential in our model. This discovery demonstrates that the patient’s comprehension of the health risks is important for reducing no-shows.

**Conclusion**

We employed state-of-the-art machine learning and causal inference methods to study no-shows in the data of more than 160,000 screening and diagnostic breast imaging appointments made during 2020. Our results suggest that a patient’s perceived risk for breast cancer, as implied by the diagnostic phase, is the most influential factor in their decision to attend an imaging appointment. Additional novel predictors correspond to time-based factors indicating the growth of COVID-19 morbidity. After adjusting for confounding biases with the AdvBal method, the effect of closures on the no-show rate largely diminished. This was especially noted in patients in advanced diagnostic phases, who seemed to maintain their lower rate of no-shows. On the other hand, the impact of closures was most pronounced in patients at age 60+, presumably due to their higher risk for coronavirus complications. Our study further demonstrates the need for personalized medicine to target additional women at higher risk for breast cancer within subpopulations associated with increased no-show rates. In a future study, we intend to leverage richer clinical data to estimate the coronavirus and cancer risks more accurately and analyze them in the context of no-shows.

**Acknowledgements**

This study was partially supported by the Israel National Institute for Health Policy Research.

**References**

Discovering Associations between Social Determinants and Health Outcomes: Merging Knowledge Graphs from Literature and Electronic Health Data

Yoonyoung Park¹, Natasha Mulligan², Martin Gleize², Morten Kristiansen³, Joao H Bettencourt-Silva²
¹IBM Research, Cambridge, MA, USA; ²IBM Research Europe, Dublin, Ireland; ³IBM Watson Health, Dublin, Ireland

Abstract

Social Determinants of Health (SDoH) are an increasingly important part of the broader research and public health efforts in understanding individuals’ physical and mental well-being. Despite this, non-clinical factors affecting health are poorly recorded in electronic health databases and techniques to study how SDoH might relate to population outcomes are lacking. This paper proposes an approach to systematically identify and quantify associations between SDoH and health-related outcomes in a specific cohort of people by (1) leveraging published evidence from literature to build a knowledge graph of health and social factor associations and (2) analysing a large dataset of claims and medical records where those associations may be found. This work demonstrates how the proposed approach could be used to generate hypotheses and inform further research on SDoH in a data-driven manner.

Introduction

With the increasing emphasis on delivering high value and accessible care, there is a growing awareness of the importance of non-clinical factors contributing to people’s physical and mental well-being collectively called social determinants of health (SDoH). A large body of prior work has shown significant effects of SDoH on health related outcomes. For example, measures of SDoH were shown to be associated with increased risk of preterm birth, readmission risk, hospitalization rate, or healthcare utilization. The global SARS-COV-2 pandemic has further revealed the stark inequity and inequality in healthcare resource across the US, often associated with different dimensions of SDoH such as race, income, education level, or job security. Timely and effective intervention regarding SDoH seems more important than ever.

From a healthcare provider or policy maker point of view, having insights on social factors that affect his/her population of interest can help design monitoring and public health interventions, prioritize resource allocation, and achieve health equity. Limited insights on the prevalent social issues of the population may exist, but it’s less likely that a thorough understanding of how much a particular issue affects which subset of the population exists. For example, a hospital knows that a large proportion of the population in its catchment area suffers from poor transportation, but is unclear about how the transportation problem adversely impacts different aspects of population health.

The interest in SDoH has led to a need for analytic and data-driven solutions, with a newly coined term social informatics. Probably the biggest challenge pointed out by both researchers and practitioners is the lack of data on SDoH. Existing health databases such as electronic health records or administrative claims data often do not collect information on SDoH in a reliable manner. A new set of codes was introduced during the conversion from ICD-9 to ICD-10 to better capture SDoH in claims data, but the utility of these Z-codes known to be under utilized are not fully understood. Area-based composite indices of SDoH based on US Census data have become more popular as a public resource. However, limited applications exist to date among healthcare organizations possibly due to both technical and privacy related concerns. Research on the impact of social factors is active in domains other than healthcare as well and vast amount of literature exist, but incrementally digesting that information and producing actionable insights would not be a feasible or efficient task. Therefore, developing tools to utilize previously untapped sources of data to guide clinical and policy decision making would be highly beneficial from both research and practice point of views.

In this work, we introduce a systematic approach to identify and quantify associations between SDoH and health related outcomes in a specific cohort of people. Our approach has two parallel pipelines, one that mines associations from published evidence in PubMed, and the second which identifies related cohorts in electronic health data. The combined results then inform the subsequent analysis with stronger hypotheses on the association between SDoH
of interest and specific health-related outcomes for further investigation. The main contributions of this work are the construction of a PubMed knowledge graph for identifying SDoH associations based on our prior work and the demonstration of methods to augment it with a real-world health data set.

Background and Related Work

Social factors are reported to account for more than half of the deaths in the US in any given year. As the focus on population health grows, interest is also rising in data and analytic solutions to better understand SDoH and how we might integrate them with clinical outcome information. One way to achieve this is to analyze electronic health data based on a priori hypotheses, such as administrative claims data generated from insurance processes or electronic medical records (EMRs) generated by providers. However, these data are built for purposes other than research and do not suitably capture information on patients’ social factors. Despite the efforts to address this limitation by, for example, introducing Z-codes in ICD-10, the utilization level of Z-codes has been very low in Medicare beneficiaries. Unstructured data, such as case notes, may be an additional source of SDoH information when coding is unavailable or sparingly documented. Recent efforts are also underway to create or improve existing standards and represent SDoH information, such as the SIREN and Gravity Projects, LOINC’s models for the representation of screening assessments and measures of SDoH, or HL7 FHIR profiles and extensions, among others.

Knowledge extraction and text mining techniques may be used to discover associations and patterns in large datasets. The application of data and text mining in health informatics is increasing especially with the availability of large electronic health data and improved computing power. Recently, researchers are realizing the potential of text mining through Natural Language Processing (NLP) in medicine to gain additional insights from traditionally underused unstructured data. For example, text mining has been used to extract medical concepts or SDoH related information from patients’ medical records. Previous work has also focused on extracting homelessness and adverse childhood experiences large corpora of clinical notes while other work has used ontology-driven tools embedded in health records to identify individuals at an increased psychosocial risk.

Our prior work in this domain is probably one of the first attempts to utilize a novel source of textual data, PubMed, for SDoH research. Published peer-reviewed articles available in PubMed provide a great opportunity to gather new insights from a very large corpus of data. In this paper, we build on the prior work and extend the approach to identify and quantify the relations between SDoH and health outcomes in a real-world cohort through electronic health data analysis.

Methods

Figure 1: Illustration of the approach used in this paper
Figure 1 illustrates our approach to assessing the associations with SDoH using two sources of data. We begin by imagining the hospital decision maker mentioned earlier, who has some knowledge about SDoH challenges in her population and wishes to know more about their impact on an undetermined set of health related outcomes. Starting from the specific social factor of interest, we identify possible associations by creating a knowledge graph (KG) based on published abstracts in PubMed. By doing this, we are collecting all available insights on the chosen SDoH topics from all research publications. However, information extraction based on co-occurrence gives at best measures of correlation which may or may not be applicable for the specific target population. While they are useful for hypothesis generating purposes, this is a critical limitation for the decision maker because information extraction from such sizable corpus will return numerous pairs of correlated terms with weak evidence. We augment the findings from PubMed by adding electronic health data, which represents data collected from the population of interest. Cohort-specific data analysis may seem self-sufficient to understand the SDoH effect, but in reality, electronic health data poorly captures SDoH and even when these are recorded, some or all of the results may be spurious correlations due to nonrandom data collection and other unobserved confounders. In addition, electronic health data are highly fragmented, often unable to produce long-term effect estimations. We therefore focused on the intersection of the two KGs that generate stronger hypotheses informing the subsequent cohort analysis, which involves quantification of the strength of association for specific SDoH-outcome pairs to generate actionable insights.

SDoH for Information Extraction and Data Analysis

We focused on two specific SDoH factors, housing problems and unemployment. The motivation for choosing housing problems is largely based on the prior evidence from Centers for Medicare Medicaid Services (CMS) showing that the most utilized Z-code among Medicare beneficiaries was Homelessness, in addition to other studies reporting adverse effects of unstable housing issue on health outcomes. Ensuring we have a reasonable prevalence of the ICD social codes in the Claims-EMR Data was important to reliably quantify the association level for these codes. The motivation for choosing unemployment came from our prior work monitoring Google Trends across several SDoH dimensions in the year 2020. The term ‘Unemployment’ had the most significant increase in terms of the average interest compared with previous years, implying that employment was a severely affected SDoH dimension at the outset of the global pandemic. Although it is not as frequently recorded in data as housing problems, we believe examining unemployment would be a timely and informative analysis in light of the SARS-COV-2 pandemic, as shown in recent studies.

Information Extraction from PubMed

![Concept network representing housing problems. Each hexagon is a category and circles are UMLS entities that represent each category. Circles that are grayed out were not found by PubMed MetaMap.](image-url)
A natural way to mine relationships between socio-medical concepts is to look for their co-occurrence in published literature\(^1\). Although a high frequency of co-occurrence does not directly indicate association between concepts in the real world, knowing that they are often talked about together is a good indication that the link has been explored by the scientific community at the very least.

We indexed the full 2019 MEDLINE/PubMed Baseline\(^2\) which notably includes the abstracts of research articles. We used MetaMap\(^3\) to tokenize and identify UMLS concepts in the sentences of the abstracts, and indexed each single sentence with Lucene\(^4\) so that it could be retrieved using multiple annotation layers, like words and phrases, UMLS semantic types, or UMLS concepts. It is possible to apply statistical methods on medical text alone but it is generally preferred to overlay a text corpus with concept annotations from an ontology like UMLS to address language variation: there is more than one way to designate the same entity. UMLS is a very extensive ontology\(^5\) and the concepts identified by MetaMap vary widely in nature, so we restricted the medical concepts to only semantic type\(^6\) representing health issues that individual people might have, like: Disease or Syndrome, Sign or Symptom. UMLS semantic types are an upper-level ontology on top of UMLS, providing broader categories that each UMLS concept can fit into\(^7\). In addition to this subset of UMLS concepts, we also highlighted Social Determinants of Health (SDoH) annotations in our indexed version of PubMed. As a starting point, we used a SDoH lexicon of concepts defined in-house and strongly inspired by the World Health Organization’s definition\(^8\). Different coding terminologies represent SDoH in different ways with varying degrees of detail and specificity. Our team then semi-automatically mapped the SDoH concepts to UMLS concepts: looking manually for the closest denomination and checking that queries using this UMLS concept yielded a non-negligible amount of results in PubMed. The mapping process itself is a challenging task, with some identified SDoH not mapping to any single UMLS concept: for example “Teenage Pregnancy” could require a combination of concepts to properly describe both the teenage and pregnancy aspects of the concept. The SDoH looked at in this contribution, identified in the previous section – Housing and Unemployment – could be straightforwardly mapped to the corresponding UMLS concept. Figure 2\(^9\) shows the SDoH Housing and related subterms, mapped to UMLS concepts.

In our experiments, we queried the index for any pair of a SDoH Housing or Unemployment, and another SDoH or another UMLS concept – as annotated by MetaMap and restricted as described above. Since our index is a sentence-level index, this defined the set of sentences where the SDoH appears together with another socio-medical concept of interest. It is important at this point to note that MetaMap’s coverage is not perfect and some concepts in UMLS are never detected by it – even on such a large collection of documents as PubMed. In Figure 2\(^9\) MetaMap detected no instance of the grayed out concepts in PubMed despite them being defined as children of Homelessness in UMLS. Grayed out concepts could be eligible for further revisions of MetaMap’s annotation of PubMed as it is important to continuously improve coverage through time.

To build a knowledge graph, we created edges between concepts found together in at least one sentence, and weighted these edges using relative frequency. Formally, the edge \((x, y)\) defined between concept \(x\) and concept \(y\) is weighted with \(W(x, y) = P(y | x)\), a conditional probability estimated on our corpus by dividing the size of the set of sentences containing both \(x\) and \(y\) (called cofrequency) by the overall frequency of \(x\) in the corpus. An optional step in the graph building process is to prune edges with too low of a weight, to get rid of noisy relations. In this contribution however, this is not necessary as our conclusions will focus on the top ranked relations.

**Analysis of Claims-EMR Data**

We used IBM\(^{®}\) MarketScan\(^{®}\) Explorys\(^{®}\) Claims-EMR Data (CED), created by linking administrative claims (IBM\(^{®}\) MarketScan\(^{®}\) Research Databases) and electronic medical records (IBM\(^{®}\) Explorys\(^{®}\) EMRs)\(^{10}\) The claims data comes from both privately insured individuals through a variety of fee-for-service, fully and partially capitated plans and individuals with employer-based Medicare supplemental insurance. The EMR data provide additional details. The linked CED is a statistically de-identified, standardized, and normalized data set that contains 5 million patient-level records on demographics, diagnostic and procedure codes, lab tests and vital signs, admission records, payments, and

---

4. https://www.who.int/health-topics/social-determinants-of-health
prescription drug information. Social aspects of patients’ lives are captured by the V-codes in ICD9 and Z-codes in ICD10 in electronic health data, and both were used to identify recorded SDoH issues in patients’ claims. Based on co-occurrence of codes, we extracted correlated diagnoses in the cohort of patients with that SDoH. For housing problems, we combined all housing related codes including Homelessness, Inadequate housing, Other specified housing or economic circumstances, etc. The ICD codes were grouped to represent more general diagnosis groups (Table 3) when summarizing and interpreting the results. By using the cohort created from all available CED data, we then create a very general cohort of people residing and receiving healthcare in the US (Table 1) - however, this can be replaced by any specific cohort data based on study settings.

Table 1: Demographic characteristics of cohorts created from CED data

<table>
<thead>
<tr>
<th></th>
<th>All Population</th>
<th>Housing Cohort</th>
<th>Unemployment Cohort</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (Mean, Std)</td>
<td>43.6 (21.1)</td>
<td>51.7 (19.9)</td>
<td>43.5 (13.3)</td>
</tr>
<tr>
<td>Female (%)</td>
<td>54.0%</td>
<td>55.2%</td>
<td>56.7%</td>
</tr>
<tr>
<td>Insurance - Medicare</td>
<td>32.6%</td>
<td>37.7%</td>
<td>3.6%</td>
</tr>
<tr>
<td>Insurance - Commercial</td>
<td>87.4%</td>
<td>82.3%</td>
<td>96.4%</td>
</tr>
<tr>
<td>Years of enrollment (Mean, Std)</td>
<td>5.3 (4.2)</td>
<td>4.6 (4.0)</td>
<td>5.1 (4.1)</td>
</tr>
</tbody>
</table>

Results

Several relations were found from PubMed under Disease or Syndrome and Sign or Symptom types (Table 2). For housing problems, the majority of the disease terms were associated with infectious diseases (Infection, Tuberculosis, Malaria, Mastitis, Parasite Infection, or Worms) or chronic conditions (Asthma, Obesity, AIDS, Diabetes, Stroke, and Respiratory disease). A number of mental health related symptoms were identified such as Emotional Depression,
Mental symptoms, or Acute stress. There were also a number of irrelevant or erroneous terms found as expected. Notably, common English words like fed, ten or march are annotated incorrectly as acronyms by MetaMap (e.g. FED for Fish Eye Disease) and their frequencies are skewed as a result. Similarly for unemployment, a diverse range of diagnosis and symptoms was extracted. Some degree of overlap with housing problems is observed but with less emphasis on infection-related diagnoses. These associated terms from a large pool of published research can generate a number of hypotheses regarding the relationship between SDoH and clinical conditions. For example, one can hypothesize that housing problems, such as homelessness, may expose a person to environments with higher probabilities of microbial infection; or that unemployment can put a person under severe mental stress that results in insomnia. One can also think about less obvious but inferred hypotheses - living conditions with housing insecurity may lead women to suffer from mastitis with little postpartum support; being unemployed can delay dental visits and lead to more people having caries. But these hypotheses are not yet specific to the population of interest and may not be applicable.

### Table 2: Top 20 Diseases or Syndromes (dsyn) and Signs or Symptoms (sosy) associated with Housing Problems (C0014003) and Unemployment (C0014003) in PubMed.

We then obtained the top 10 most prevalent diagnoses associated with either housing problems or unemployment from an analysis of CED data (Table 3). Unsurprisingly, many were chronic conditions known to be highly prevalent among general population such as hypertension, diabetes, hyperlipidemia, or reflux disease. However, an interesting observation was that the two SDoH generated the same list of top 10 diagnosis with different order of prevalence. Compared to housing problems cohort in which chronic physical conditions were the top 3 most prevalent, psychiatric diagnosis such as mood disorders or anxiety took place in the top 3 diagnoses in the unemployment cohort. The nature
of health related challenges associated with a social factor may differ across different SDoH.

<table>
<thead>
<tr>
<th>Housing Problems</th>
<th>#</th>
<th>Disease</th>
<th>ICD-9</th>
<th>ICD-10</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>Essential hypertension</td>
<td>401.x</td>
<td>110</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>Disorders of lipid metabolism (including Hyperlipidemia)</td>
<td>272.x</td>
<td>E78.x</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>Esophageal reflux (GERD)</td>
<td>530.81</td>
<td>K21.9</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>Overweight and obesity</td>
<td>278.x</td>
<td>E66.x</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>Anxiety disorder</td>
<td>300, 300.00</td>
<td>F40.x-F41.x</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>Episodic mood disorder (including depression)</td>
<td>296.x, 311</td>
<td>F30.x-F39.x</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>Drug abuse including tobacco and alcohol</td>
<td>305.x</td>
<td>F10.x-F19.x</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>Diabetes</td>
<td>250.x</td>
<td>E08.x-E13.x</td>
</tr>
<tr>
<td>9</td>
<td></td>
<td>Vitamin D deficiency</td>
<td>268.x</td>
<td>E55.x</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td>Chest pain</td>
<td>786.5x</td>
<td>R07.89, R07.9</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Unemployment</th>
<th>#</th>
<th>Disease</th>
<th>ICD-9</th>
<th>ICD-10</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>Episodic mood disorder (including depression)</td>
<td>296.x, 311</td>
<td>F30.x-F39.x</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>Essential hypertension</td>
<td>401.x</td>
<td>110</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>Anxiety disorder</td>
<td>300, 300.00</td>
<td>F40.x-F41.x</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>Drug abuse including tobacco and alcohol</td>
<td>305.x</td>
<td>F10.x-F19.x</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>Diabetes</td>
<td>250.x</td>
<td>E08.x-E13.x</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>Overweight and obesity</td>
<td>278.x</td>
<td>E66.x</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>Esophageal reflux (GERD)</td>
<td>530.81</td>
<td>K21.9</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>Disorders of lipid metabolism (including Hyperlipidemia)</td>
<td>272.x</td>
<td>E78.x</td>
</tr>
<tr>
<td>9</td>
<td></td>
<td>Chest pain</td>
<td>786.5x</td>
<td>R07.89, R07.9</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td>Vitamin D deficiency</td>
<td>268.x</td>
<td>E55.x</td>
</tr>
</tbody>
</table>

Table 3: Top 10 most prevalent diseases associated with Housing problems and Unemployment in the Claims-EMR dataset.

Combining the two KGs, represented with the extracted relations, led to the identification of several concepts that are directly or indirectly linked to the concepts in the other KG. Disease concepts such as diabetes, obesity, and hypertension appear in both PubMed and CED results. Concepts from PubMed like emotional depression or acute stress can be linked to mood disorder diagnosis in CED. We calculated the prevalence of the top 10 diseases in the general CED population and in the cohort specified by each of the two SDoH examined. The largest relative difference in prevalence was observed for diabetes and episodic mood disorder (Table 4) so we focused on these two in the subsequent analysis.

The last part of our experiment was quantifying the level of association in the CED cohort. Specifically, we examined whether the healthcare cost and utilization level among those with a disease diagnosis differ across the general cohort and the subgroups of people with housing problems or unemployment codes. As Table 5 shows, for patients with diabetes diagnosis, having a code for housing problems was associated with greater outpatient cost (per year per patient) and outpatient service use measured by number of claims (per year per patient). Similarly, for patients with mood disorders diagnosis, having a code for housing problems was associated with greater outpatient cost and outpatient service use. Notably, emergency department cost was significantly higher for patients with housing problems or unemployment code compared to the general population, for both diabetic and mood disorder patients. On the other hand, we observed that inpatient service utilization and related cost was higher for the general CED population compared to the SDoH-specific subgroups.

Discussion and Conclusion

We describe a systematic approach to identify and quantify meaningful associations between SDoH and health related outcomes using published knowledge as well as electronic health data. The novelty of this work is the augmentation of knowledge extraction from PubMed by adding a cohort-specific, independent source of clinical data. The final set of pruned associations with the chosen SDoH can be informative and actionable from a decision maker point of view – whether it be a population health specialist in a hospital or a policy maker at a public health department – because it provides augmented insights specific to the population group of interest that may not be observable in data alone. As
our approach does not attempt to identify causal relationship, evaluating the impact of SDoH should be followed by additional data collection or more rigorous analysis. Ongoing efforts to address the need for a more comprehensive SDoH terminology, such as methodologies to detect new SDoH concepts or the Gravity project should, in the future, provide additional structure to the work presented in this paper.

The ability of our proposed approach to identify meaningful SDoH associations from KGs were supported by literature findings. The increased emergency service utilization among people with housing problems we observed in the data has also been reported in prior literature. The observation that unemployment is associated with higher prevalence of mental health diagnosis can also be supported through prior evidence. The validation of selected hypotheses through literature means that other associations we observe in data are potentially worth further investigating. For example, one relation we observe from PubMed knowledge extraction is between housing problems and respiratory diseases or symptoms, which is not captured by the 10 most prevalent ICD codes in the CED data. While it is possible that this is just a spurious association, it can also be a meaningful association not captured due to incomplete data collection or low prevalence. Considering the disease burden and clinical significance of managing chronic respiratory diseases, it may be worth investigating this association further to see whether this association is true, and if so what may be the factors that contribute to both housing problems and diseases like asthma or COPD, such as geographic characteristics.

There are extended application use cases for this approach. For example, housing problems may be selected as an input for outcome risk modelling to adjust for confounding. Also, it can be used as an input for predictions to improve the model performance.

The known limitations of observational health data are applicable for this study, including non-randomly missing diagnosis, incomplete capture of data such as lack of information on the duration of illness or SDoH, and coding errors or variability in coding practice. A claims data point is generated at the time of a patient’s interaction with health system, so those who lack access to care, who will likely have more severe SDoH problems, may be absent from the data. By combining claims with the EMR data, information from CED can provide more complete picture of individual patients and increases the likelihood of capturing all data compared to using only claims.

References


Implementation matters: How patient experiences differ when genetic counseling accompanies the return of genetic variants of uncertain significance

Harsh V. Patel, BS1, Nora B. Henrikson, PhD MPH2, James D Ralston, MD MPH2, Kathleen Leppig, MD3, Aaron Scrol, BA2, Gail P. Jarvik, MD PhD1, Shannon DeVange, MS CGC1, Eric B Larson, MD MPH2, Andrea L. Hartzler, PhD1

1University of Washington, Seattle Washington; 2Kaiser Permanente Washington Health Research Institute, Seattle Washington; 3Kaiser Permanente Washington, Seattle Washington

Abstract

Precision medicine presents challenges for effective return of results (ROR) to patients, particularly for variants of uncertain significance (VUS) where the need for genetic counseling and the impact of results are underexplored. We investigated patients’ experiences with VUS ROR. Through interviews we compared experiences of patients who were referred to genetic counseling with those not referred. Although participants from both groups (n=16) reported curious enthusiasm and relief after ROR, the 5 referred participants reported less confusion, less disappointment, and better confidence in understanding their results than the 11 non-referred participants. Although VUS did not impact healthcare or daily lives, some participants who shared VUS fostered communication about future healthcare. Suggested ROR improvements included patient-friendly terminology, on-demand education, and ongoing consultation. Although patient experience of VUS improved when ROR involved expert consultation, scarcity of genetic counselors presents challenges. Improving the ROR process with patient-centered solutions could enhance the patient experience of receiving VUS.

Introduction

As precision medicine becomes integrated into healthcare, questions emerge about how best to return genetic results to patients, particularly for variants of uncertain significance (VUS). VUS are genetic changes for which the association with disease is unknown. Although the American College of Medical Genetics and Genomics recommends reporting incidental or secondary findings from actionable genes that are not the target of testing in clinical practice,1,2 they recommend against returning VUS as a secondary finding. However, when VUS are primary findings, detected for a gene that is the target of the ordered test, VUS are generally clinically returned even though most are not clinically actionable nor found to be pathogenic.3

Strategies for returning genomic results range from in-person/phone consultation4,5 to passive notification through patient portals or mailed letters.6 Prior work has examined patient experience with return of genomic results in general. For example, patient-friendly genomic test reports have been designed to improve patient engagement and understanding of complex genetic data.7 More advanced tools for returning results, such as “My46”, offer self-guided management of results.8 Other researchers have infused direct-to-consumer personal genomic reports interactive features to enhance understanding.9 Prior work has examined patient experience with such patient-facing strategies for return of genomic test results. For example, patient-facing genomic reports have been shown to improve patient communication with providers, educators, and therapists, which led to increased engagement and satisfaction.10 Yet, how such strategies impact the patient experience of receiving VUS remains poorly understood.

Clift et al.11 call attention to the potential for patient misinterpretation of VUS and the need for counseling and education for both patients and providers. Although the behavioral and experiential consequences of receiving a VUS results are not well studied, early evidence suggests taking caution in how VUS are returned. In a study on return of Lynch Syndrome related VUS results, Soloman et al.12 found that patients may be surprised by VUS and interpret its clinical significance in a wide range of ways. Similarly, patients who received VUS demonstrated mixed understanding and expressed both uncertainty about the impact of VUS on clinical management and concern for family members’ wellbeing.13 However, these studies did not examine the impact of how results were returned on patients’ experiences. Given the uncertainty associated with VUS, better understanding this experience can inform effective return of result (ROR) strategies that patients find acceptable.
Methods

The objective of this qualitative study was to investigate the ROR experience of patients for VUS. We conducted semi-structured interviews to compare participants’ experiences of receiving VUS with and without referral to clinical genetics.

Study setting and recruitment

This study takes place within the context of The Electronic Medical Records and Genomics network (eMERGE).\textsuperscript{14} eMERGE began in 2007 to develop best practices for genomic research in bio-repositories linked with electronic health records (EHR). As a collaborating site in eMERGE, Kaiser Permanente Washington and the University of Washington conducted a comprehensive program of genomic discovery and clinical implementation research\textsuperscript{15} in which genetic results for bio-repository participants were integrated into the EHR.

Our study population was eMERGE participants with a colorectal cancer (CRC) diagnosis or colon polyps and a VUS, including those referred and not referred to clinical genetics to receive counseling for their VUS. Only participants with a VUS in one of the genes associated with Lynch syndrome were referred for genetic counseling and in-person ROR. Other VUS were not offered genetic counseling (Figure 1). Inclusion criteria were being a Kaiser Permanente member and remembering receiving the VUS.

Participants with a VUS who were referred to genetic counseling received a mailed letter describing the study and results written at 8th grade level and a referral for clinical genetics consultation. Results were returned during consultation, and then placed in the EHR and accessible to the participant via the patient portal. The participant’s primary care provider was sent a copy of the results and consultation. Participants with a VUS who were not referred also received the mailed letter, and their results were sent to their primary care providers and placed in the EHR, where they were accessible to the patient via the patient portal. Participants who were not referred were considered an appropriate comparison as they did not have a VUS in one of the Lynch syndrome genes and as a result did not receive in-person genetic consultation.

We sent study invitation letters to 108 participants of whom 32 were referred (30%) and 76 were not referred (70%). Forty-six (43%) responded (17 referred, 29 not referred) and completed a phone screening. Of those screened, 29 declined (11 referred, 18 not referred) and one who was referred failed to meet inclusion criteria.

Data collection and analysis

From March to April of 2019, one member of the research team (AH) conducted interviews using a semi-structured guide covering participants’ experience with the VUS ROR process (Table 1). Interview questions covered the following domains: 1) emotional and cognitive reactions to receiving VUS, 2) whether and how they shared results with healthcare providers, family, or friends, 3) perceived impact of results on their healthcare and daily life, and 4) recommended improvements to the ROR process given perceived benefits and barriers. We collected age, gender, race, ethnicity, education, and self-reported familiarity with genetics on a 5-point Likert scale from 1 “not at all familiar” to 5 “extremely familiar”.

Interviews were audio-recorded and transcribed for qualitative analysis. Two coders (HP, AH) used template analysis\textsuperscript{16} to deductively code transcripts in Dedoose\textsuperscript{17} following the four domains of the interview guide. Once the coders achieved intercoder reliability of K=0.90 on a 25% sample of transcript excerpts, the remaining transcripts were split between coders and independently coded. After coding was complete, we compared experiences by domain between participants who were referred and not referred.

Results

Participants

Of the 16 participants (P1-P16), 5 were referred for genetic counseling (P5, P6, P8, P10, P12) and 11 were not referred. Participants were largely white, non-Hispanic/Latino, and female, ranged in age from 43 to 82 years old (mean=66), and had varied education and familiarity with genetics (Table 2).
Figure 1: Colorectal cancer VUS ROR workflow detailing how participants were assigned referred/non-referred status after review by Kaiser Permanente Washington (KPWA) geneticist and how results were sent to patients.

Legend
ROR: Return of Results, KPWA: Kaiser Permanente Washington, VUS: Variant of Uncertain Significance, CRC: Colorectal Cancer, PCP: Primary Care Provider, NOK: Next of Kin, LP: Likely Pathogenic, P: Pathogenic, EMR: Electronic Medical Record, CDS: Clinical Decision Support

CRC genes with results that were returned: APC, BMPR1A, JAK2, MLH1, MSH2, MSH6, MUTYH, POLE, PTEN, SMAD4, STK11, PMS2
Table 1. Interview guide

<table>
<thead>
<tr>
<th>Domain</th>
<th>Prompt</th>
</tr>
</thead>
</table>
| Reaction to VUS | • What do you remember about receiving your genetic test results? How did you receive them?  
• How did you feel about receiving your results (i.e. emotional reaction)  
• What concerns or questions did you have about your results? What, if any information was missing? What do you think this result means for your future risk of getting colorectal cancer?  
• Overall, how would you describe your experience with receiving your results? |
| Sharing VUS | Did you discuss or share your results with anyone?  
• a healthcare provider? If so, who, when, and what did you discuss?  
• a family member or friend? Without telling me any personal details about your family member, can you describe who, when and what you discussed? |
| Perceived impact | • Thinking back, what do you see as positives or negatives about having your genetic testing results returned to you?  
• How have your results influenced your healthcare, if at all?  
• How have your results influenced your day-to-day life, if at all?  
• How have your results influenced the lives of family members, if at all? |
| Recommended improvements | • If could do this process over again, would you choose to receive your results? Why, or why not?  
• What went well about the process? (how notified, educational needs, provider support)  
• What would you change to improve the process? |

Table 2. Participant Demographics

<table>
<thead>
<tr>
<th></th>
<th>All N (%)</th>
<th>Referred N (%)</th>
<th>Not referred N (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean age (SD)</td>
<td>66 (11) years</td>
<td>72 (7) years</td>
<td>63 (11) years</td>
</tr>
<tr>
<td>Median (range)</td>
<td>69 (43-82) years</td>
<td>73 (63-82) years</td>
<td>68 (43-75) years</td>
</tr>
<tr>
<td><strong>Sex</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>5 (31%)</td>
<td>2 (40%)</td>
<td>3 (27%)</td>
</tr>
<tr>
<td>Female</td>
<td>11 (69%)</td>
<td>3 (60%)</td>
<td>8 (73%)</td>
</tr>
<tr>
<td><strong>Race</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>11 (69%)</td>
<td>4 (80%)</td>
<td>7 (64%)</td>
</tr>
<tr>
<td>African-American</td>
<td>1 (6%)</td>
<td>-</td>
<td>1 (9%)</td>
</tr>
<tr>
<td>More than 1 race</td>
<td>1 (6%)</td>
<td>1 (20%)</td>
<td>-</td>
</tr>
<tr>
<td>Undisclosed</td>
<td>3 (19%)</td>
<td>-</td>
<td>3 (27%)</td>
</tr>
<tr>
<td><strong>Ethnicity</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hispanic or Latino</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Not Hispanic or Latino</td>
<td>15 (94%)</td>
<td>5 (100%)</td>
<td>8 (73%)</td>
</tr>
<tr>
<td>Undisclosed</td>
<td>1 (6%)</td>
<td>-</td>
<td>3 (27%)</td>
</tr>
<tr>
<td><strong>Education level</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High school degree</td>
<td>3 (19%)</td>
<td>-</td>
<td>1 (9%)</td>
</tr>
<tr>
<td>College degree</td>
<td>7 (44%)</td>
<td>2 (40%)</td>
<td>4 (37%)</td>
</tr>
<tr>
<td>Post graduate degree</td>
<td>6 (37%)</td>
<td>3 (60%)</td>
<td>3 (27%)</td>
</tr>
<tr>
<td>Undisclosed</td>
<td>-</td>
<td>-</td>
<td>3 (27%)</td>
</tr>
<tr>
<td><strong>Familiarity with genetics (out of 5)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Median (range)</td>
<td>2 (1-4)</td>
<td>3 (2-4)</td>
<td>2 (1-4)</td>
</tr>
<tr>
<td><strong>CRC status</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Colon polyp</td>
<td>12 (75%)</td>
<td>3 (60%)</td>
<td>9 (82%)</td>
</tr>
<tr>
<td>CRC diagnosis</td>
<td>4 (25%)</td>
<td>2 (40%)</td>
<td>2 (18%)</td>
</tr>
</tbody>
</table>
1. Reaction to VUS

Participants’ emotional reactions varied when asked “Overall, how would you describe your experience with receiving your results?”. Those who were referred primarily reported enthusiasm, curiosity, and relief:

“I like that kind of stuff so I was excited to get it. Didn't really understand what it was.” (P6, referred)

“I felt curious and intrigued … It was something that I was kind of interested in, seeing how it happened and what the results were.” (P12, referred)

“I was a little bit relieved, I guess, because from the day I had the colonoscopy and they took the sample, I -- the surgeon said, you know, I saved your life. And then -- but everybody else since then has been very noncommittal as to whether I had cancer or not. And so finally, this study, I will say that, that it came right out in front and said it was cancerous.” (P5, referred)

Although most who were not referred shared similar sentiments of curious enthusiasm and relief, a few expressed confusion and disappointment:

“Well, it was confusing. They said I had a marker … but the implication was that I had cancer.” (P1, not referred)

“I felt a little concerned because they still -- I mean, I just remember they weren't being specific … I haven't really been told exactly what they have found … So I'm still a little confused.” (P4, not referred)

When asked “How well did you understand your results?”, most participants expressed challenges. Six of the 11 participants who were not referred indicated they did not understand the results on their own:

“I was excited about taking the test, but I don't remember understanding the results … I know it was disappointing because I didn’t understand … so it didn't make, you know, much sense to me.” (P2, not referred)

“I thought it was a little bit too much in medical terms … and it was kind of hard to understand.” (P13, not referred)

“It was a lot of inconclusive stuff, and I don't remember learning a lot.” (P15, not referred)

In contrast, referred participants reported feeling more confident in understanding the results after speaking with the genetic counselor:

“[I] didn't really understand what it was. I went in for an interview [genetic counseling] which helped more … I sat down one on one … and went through, you know, I think there was a list of genes and things that impacted me … It was very informative.” (P6, referred)

“You really definitely need to have the one-on-one with a geneticist or someone in the genetics department to explain the technical sides… There has to be that discussion.” (P8, referred)

“We were able to sit down with the person … she walked us through it … It was clear. Both my wife and I understood it clearly.” (P10, referred)

2. Sharing VUS

Most participants (13/16) shared their results through email or conversation with family members, primarily children, spouses, or siblings. Most participants with colon polyps (11/12) shared with family whereas two of the four with CRC did so.

Most participants who shared with family, both those who were referred and those who were not referred, expressed a duty to share VUS with family in case of future health implications:

“But the fact that I did, in fact, have a marker was something that I wanted to let my family members know. ... They need to be aware of that. They needed to take action and that needed to be part of their health strategies. ... I remember being very clear with my nieces and with my brother.” (P10, referred)

“I sat him [son] down and said, do you realize that this could be a possibility, you know, this really could, and it's important that find out as much as you can about this, whether genetically you would be at risk.” (P3, not referred)

“I just emailed them [my kids] that I was in this study and it looked -- and I had the marker for colon cancer
and they should put that information somewhere so they have it at the ready if they need it.” (P1, not referred)

However, other participants were selective in sharing with select family members due to inherent uncertainty in VUS:

“I talked to one of them, my sister, and said no, I'm not going to go any further than that. You know, we're all in our 70s and 80s. And there's no sense in me stirring people up. Well, yeah. Because -- because -- because the results are so iffy. So uncertain. It even says of 'uncertain significance.' That doesn't mean anything to me. So I'm not going to get people to worry about stuff they don't -- can't do anything about.” (P5, referred)

In contrast to the majority of participant who shared with family, only five participants explicitly shared with a healthcare provider. Of those five, four had colon polyps and one had a diagnosis of CRC. For example, one of the two referred participants who shared with a healthcare provider made a copy of the VUS letter and gave it to them:

“[I] couldn't remember if this was passed on to my doctor or not, but I did make a copy and gave it to him ... so I figured they must know about it.” (P6, referred)

The three participants who were not referred and shared results with providers talked about the value that providers can add to help interpret and bring meaning to VUS results and their impact on future healthcare:

“Well, I asked my gastroenterologist if it [VUS] was meaningful in the plan that we had developed. 'Should I?', you know -- so we had a conversation.” (P1, not referred)

The majority of participants chose not to share their results with their healthcare providers. Some participants reached the conclusion that sharing with their provider was not warranted:

“And the first thing it [result] says is that there is a variant of unknown significance. So it may or may not mean something. We don't know if it means anything. And so -- and then reading through the report, it was clearly all scientific language. And the conclusion was you don't need to talk to your doctor about it. Nobody's really going to contact you because we don't know what this means”. (P7, not referred)

“Nobody seems to be alarmed by the information in this report, so I'm not going to let it alarm me if I have some -- one gene that looks a little weird... They're not telling me that there's any particular risk. Some unknown possible risk but we don't know, and there are a whole bunch of big words too in scientific medical language about what it is. But clearly, they don't care if I understand it, because they would have put it in different language.” (P7, not referred)

3. Perceived impact of VUS

No participants reported changes to their healthcare based on receiving VUS results. However, four participants described how receiving results reinforced healthy lifestyle choices in daily life, such as diet and cancer screening.

4. Recommended improvements

Despite barriers to understanding VUS, most participants found benefit from receiving VUS. Participants who were not referred reported increased awareness of their genetic makeup and feelings of altruism from participating in research. Participants who were referred found benefit in genetic counseling, ability to ask questions, and additional information they received beyond the mailed report.

Participants from both groups recommended patient-centered improvements to the VUS ROR process. Most participants (11/16) recommended expert consultation, links to videos, and primers on genetics and VUS:

“I do think it would be really helpful for people to get the results face-to-face with a professional who could explain what was done and what the results meant and describe the process more thoroughly if someone wanted the process described more thoroughly. I think I had a pretty good understanding of genetic testing and how that operates and what kinds of things they're discovering and all of that. But I think that rather than just getting, you know, a typical result of a test in the mail.” (P10, referred)

“I'd rather speak to somebody so I could say, well, ... what am I doing wrong? You know, if it's not genetic, then what am I doing wrong? You know? Am I overweight? Am I drinking too much? Am I not getting enough exercise? Should I eat carrots? You know, things like that. Those questions I would have liked to have asked of the people that did this type of testing” (P3, not referred)

To address drawbacks, such as feeling hindered by confusing medical jargon and inability to find clarifying
resources, several participants who were not referred suggested framing explanations in lay terminology and providing education resources:

“Well, just more in lay terms. Like on additional notes, you know, they have KC and Q1 92 percent, you know, PM as to 93, you know, that kind of thing. You have no idea what all that means...They detailed, you know, the interpretations and what they did, but I don’t -- maybe it’s just me, but I didn’t understand a lot of it. ” (P13, not referred)

“You know how, like, when you read a medical paper or a legal paper there are footnotes that you can read further about this if you go to this source. So that could be kind of helpful for people. Even if it was as simple as, you know, an article on the basics of genetic studies. ” (P16, not referred)

Participants who were referred also suggested improving after-visit documentation, more timely communication, and the potential for future consultation:

“Maybe if a CD was made of the interview at the time, that you could take with you would just accompany the written information. That way down the road you could plug it in as you're looking through the written information and have that explanation freshen you. “(P6, referred)

“I think there was a delay in the time between getting the results and having the conversation, so I think people who maybe are not as comfortable with the healthcare system and terminology might prefer a closer time opportunity to discuss the information sent to them in the mail. ” (P12, referred)

Discussion

In several ways, ROR experiences differed between participants who were referred and not referred to genetic counseling for VUS. Those who were referred reported less confusion, less disappointment, and more confidence in understanding their VUS than non-referred participants. Although some participants who were not referred also expressed positive experiences, confusion and frustration with the ROR process was evident. Several participants did not understand their results and desired expert consultation and clarifying resources. Although VUS did not appear to impact healthcare, participants who shared VUS may have fostered communication about future healthcare. All participants offered suggestions that can inform healthcare systems in patient-centered improvements to the ROR process that prioritize patient experience (e.g., patient-friendly terminology, on-demand education, ongoing consultation).

Many patients want to be included in deciding what genetic results are returned and find value in the results beyond clinical utility, yet we know comparatively little about patients’ perceptions of VUS. Although the consequences of receiving VUS are not well studied, early evidence suggests taking caution in how VUS are returned. Patients may misinterpret VUS and interpret their clinical significance diversely. For example, some women with BRCA1/2 VUS pursue mastectomy and/or salpingo-oophorectomy. Other work demonstrates similar mixed patient interpretation of uncertainty and implications of VUS. However, these studies did not examine the impact of how results were returned on patients’ experiences. Our sample may limit transferability of findings. Because participants who were referred had a VUS in one of the Lynch syndrome genes and those who were not referred did not, group differences beyond referral could have impacted results. Given limitations of our small homogeneous sample, future work whose scope is broader and examines potential racial and other disparities is needed to fully understand best-practice communication methods and gauge patient understanding. Although our small sample was from a single health system, our findings add in-depth insight for patients’ experiences into this poorly understood topic.

The process through which patients receive VUS matters - a text report alone may be insufficient and leave patients without needed assistance to interpret the results. Patients may experience less confusion, less disappointment, and greater confidence in understanding VUS when ROR involves expert consultation. Given the scarcity of genetic counselors, our findings present challenges for meeting the needs of patients in the era of precision medicine. Patient-centered solutions, such as virtual agents, educational portals, and patient-friendly formats could scale support to reach diverse audiences. However, future work should demonstrate the value and acceptability of such solutions to patients. More fundamentally, our findings give pause to whether the benefits of returning VUS outweigh the potential risks when genetics consultation is not indicated.
Conclusion

Healthcare systems should gauge the needs of patients and report genetic results in patient-friendly ways. Although genetic counselors are critical to effective ROR, information technology and processes that carefully consider patient experience could ease emerging challenges of precision medicine.

Acknowledgments

We wish to thank our participants as well as Suhk Makhnoon and Stephanie M. Fullerton who provided helpful feedback on this manuscript. This work was supported by NIH/NHGRI grants U01HG008657 and 2R01LM011563.

References


Creation of a Mapped, Machine-Readable Taxonomy to Facilitate Extraction of Social Determinants of Health Data from Electronic Health Records

Svati B. Patel, MHS MSc\textsuperscript{1} & Nam T. Nguyen, MS\textsuperscript{2}
\textsuperscript{1}Veradigm, Chicago, IL, \textsuperscript{2}Veradigm, San Francisco, CA

Abstract

A comprehensive, mapped social determinants of health (SDH) taxonomy in machine readable format was developed. The framework is intended to facilitate the extraction of social risk factors (SRFs) out of electronic health record (EHR) data and categorize them by domain and determinant to facilitate interpretation. Where other SDH frameworks have been focused on data input, this framework is designed from a data extraction point of view using EHR data in conjunction with published literature, public health policy documents, and official crosswalk maps. Frameworks developed by leading public health organizations were reviewed and synthesized to create an SDH framework comprising of 97 distinct SRFs organized under 16 domains. 2,329 medical codes across three standardized medical vocabularies, 10,896 free-text diagnosis descriptors, and 25 health insurance keywords were mapped to individual SRFs in the SDH framework. The framework is available as an open-source resource in Python dictionary or JSON format.

Introduction

Evidence on the impact of social and economic drivers on patient health outcomes has been mounting over the past two decades.\textsuperscript{1} This research posits that factors such as geography, housing, food, employment, education, and income can substantially shape health and well-being. With the push for value-based, patient-centered medicine in recent years, health care provider institutions have become increasingly focused on addressing these SDHs with the goals to improve patient outcomes and control health care costs.\textsuperscript{2,3} While much of the work and funding involved with managing social needs occurs outside the health care setting, these institutions recognize that health care providers, especially primary care, can play a pivotal role in identifying at-risk patient populations, screening for social needs, developing care plans that account for social factors, and directing patients to community and government resources.\textsuperscript{4} Implementation of a carefully considered infrastructure is required to effectively carry out these tasks. An institution’s population health informatics capabilities and its EHR are key elements of this infrastructure, particularly in identifying at-risk populations.\textsuperscript{5}

However, even identifying SDH patients presents a challenge; population health informatics professionals face a three-fold problem. First, there is no standardized definition of what factors are considered social determinants.\textsuperscript{6} Multiple public health organizations have published frameworks identifying individual SDHs and associated domains including the World Health Organization (WHO)\textsuperscript{7}, National Academy of Medicine (NAM)\textsuperscript{8}, Kaiser Family Foundation (KFF)\textsuperscript{9}, and the United States Department of Health and Human Services’ (DHHS) Healthy People 2020\textsuperscript{10}. Further, the International Classification of Diseases, Tenth Revision (ICD-10) includes a section of diagnosis codes (Z55-Z65) which provides a framework of SDH domains and social risk factors (SRFs).\textsuperscript{11} These frameworks overlap in many respects, but each also uniquely highlights important determinants and domains not addressed by the others. Without a universally agreed upon framework of SDHs and SRFs, an informatics professional must either choose one or attempt to reconcile them.

The second challenge is identifying the multitude of EHR data elements that could potentially record SRFs either directly or indirectly. Failure to carefully consider all the possible ways that SRFs could be documented in the EHR would yield only a partial picture of its prevalence in a population. EHRs in their current form contain a wealth of data elements that can be used to flag patients with social risks. Such SRFs can take the form of standardize medical vocabularies\textsuperscript{2}, unstructured text\textsuperscript{12}, demographic classifications, or administrative information.

The third challenge is the lack of a comprehensive, machine-readable map that connects data elements to a defined determinant or SRF. Creating these linkages is critical so that captured data can be aggregated, no matter how or where in the EHR the data is encoded, at a level that is meaningful from a population health perspective. The Social Interventions Research & Evaluation Network (SIREN) started the development of such a map by conducting a systematic review of standardized medical vocabularies to develop a compendium of medical codes for 20 domains and subdomains. These domains aligned with six widely recognized screening tools used to collect social and economic risk factors.\textsuperscript{13} SIREN’s compendium aligned key survey questions and patient responses to standardized codes available within the Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT), ICD-10, the
Logical Observation Identifiers Names and Codes (LOINC®), and the Current Procedural Terminology (CPT®). This pivotal work, and resulting map, facilitates documentation of SDH in clinical settings within the EHR as structured data. While the SIREN provider-focused compendium enables EHR data input, there is no complementary machine-readable tool that enables data extraction. SIREN’s aim was to be thorough in its mapping to ensure providers were able to input all potential patient responses to SDH survey instruments; it does not, however, define target SRFs for data extraction. The SIREN tool also does not address character-limited, free-text data elements or indirect data elements that could be used to identify at-risk patients.

To that end, we conducted literature reviews and utilized EHR data to derive a comprehensive, machine-readable taxonomy of SDH domains and determinants mapped to both standardized and character-limited free-text data elements. The goal was to build upon SIREN’s seminal work to create a mapped SDH taxonomy that can facilitate data extraction and serves as a tool to identify at-risk patients within an EHR. This tool, as well as other materials discussed in this paper, have been made available as an open-source resource for others in the community to use, add to, and refine.

Methods

This study involved two steps. The first step was a review and synthesis of the many SDH frameworks issued by public health organizations with the goal to consolidate them into a broader framework of domains, determinants, and SRFs. The second step focused on identification of data elements and mapping them to this SDH framework in a machine-readable format that can readily be passed into data queries. We prioritized formats that could be applied across a broad range of use cases and analytics approaches.

To aid this twofold endeavor, we introduced a unique aspect to our project: we derived and tested our SDH taxonomy using de-identified data set from three large national ambulatory EHR systems collected and maintained by Veradigm®, Veradigm, a business unit of Allscripts®, is a health information technology, analytics and intervention solutions company that manages the largest source of de-identified ambulatory patient records in the U.S. Its ambulatory EHR dataset contains medical information on patient demographics, prescriptions, problems, laboratory test results, vaccinations and allergies from providers using any of three ambulatory EHRs: Allscripts Professional EHR™ (PRO), Allscripts TouchWorks® EHR (TW), and Practice Fusion EHR (PF). Collectively, these three EHRs represent a nationally diverse range of ambulatory provider organizations, from single-provider or small group practices (PF) to mid-size physician practices (PRO) to large, single or multi-specialty physician practices (TW). By leveraging this vast and growing ambulatory EHR footprint, we were able to develop an SDH taxonomy that is generalizable and relevant to the real-world clinical setting.

Framework Development: Defining Domains and Determinants

We focused our review on commonly cited frameworks promulgated by WHO, NAM, KFF, and DHHS Healthy People 2020 as well as the one outlined in the ICD-10 section Z55 to Z65 code set (“ICD-10’s Z-codes”). The review of ICD-10 was limited to just the ICD-10’s Z-codes as this is the section of the code set that is generally recognized as containing SDH codes.15

We chose ICD-10 Z-code’s domains and determinants to be the initial outline of our SDH taxonomy. Of the five reviewed, ICD-10’s Z-codes represents the only framework with determinants that are fully articulated as SRFs. For example, while several frameworks mention literacy as a determinant under their education domain, the education domain within ICD-10’s Z-code includes the SRF illiteracy and low-level literacy. Literacy is the SDH that refers to a person’s ability to read or write. Anyone can be evaluated for literacy. Illiteracy, however, refers to a specific subgroup of people whose literacy skills are low or nonexistent. Given the value a society places on literacy skills, illiterate individuals are considered at risk. In this scenario, education is the domain, literacy is the SDH, and illiteracy is the target SRF.

The remaining four frameworks by WHO, NAM, KFF, and DHHS Healthy People 2020 were then assessed against this outline. Areas of overlap were noted. We recorded overlaps in both domains and individual SRFs. In some cases, frameworks would reference a domain such as Housing but not include individual SRFs such as Homelessness or Housing Inadequacy. Further, we also evaluated the need for reorganization of SRFs within the outline to improve conceptual groupings under each domain.

1 Codes for many other social and economic drivers can be found throughout ICD-10, but they are typically contained within sections that, unlike the ICD-10 Z-codes, do not necessarily call attention their status as SRFs and are not organized as such.
Gaps between the four frameworks and ICD-10’s Z-codes were recorded as well. We set a goal to develop an inclusive SDH taxonomy that drew upon unique aspects of the frameworks reviewed rather than a least common denominator approach. We reasoned that such a framework would help future-proof our taxonomy as the definition of SDH evolves over time. Thus, domains and SRFs that were not covered by ICD-10’s Z-codes but were mentioned in the other reviewed frameworks were evaluated for inclusion. This evaluation was conducted by both authors and decisions for adding new domains and SRFs were made on the basis of three factors: (1) the number of frameworks with coverage, (2) the utility of extracted data in health informatics and modeling, and (3) the ability for health institutions to deploy targeted inventions for the SRF.

A full copy of our comparative analysis of SDH frameworks is available online.

Data Elements: Selection and Mapping

Standardized Medical Vocabularies

The SDH taxonomy was populated with mapped ICD-10 codes, ICD-9 codes, and SNOMED codes using a number of methods. First, for areas where our two SDH frameworks overlapped, we used SIREN’s publicly available compendium\(^1\) of mapped ICD-10 and SNOMED codes to identify codes to add to the SDH taxonomy. SIREN’s compendium did require modification to accommodate our data extraction-oriented taxonomy. We manually distilled SIREN’s mapped standardized medical vocabularies down to just the codes representing SRFs. For example, SIREN’s Income/Poverty determinant was mapped to SNOMED codes that described a range of income levels, from wealthy to middle-class to destitution. For our taxonomy, however, we only included the codes associated with the SRFs low income, destitution, and poverty. We also eliminated duplicative mappings to ensure that each code was assigned to only one determinant. The ICD-10 code Z60.8 (other problems related to social environment), as an example, was mapped to five different social determinants. Such multiple mappings would make data aggregation difficult and distort comparative results. In these situations, we mapped the code to the most appropriate determinant based on official descriptors and public coding guidelines.

Literature reviews\(^1\), value sets\(^18\)\(^20\)\(^23\), and keyword searches of official code descriptors further supplemented our code search and mapping activities. We also obtained medical codes that were collected as part of our free-text diagnoses descriptors search strategy which is described in greater detail in the next section.

Lastly, we conducted a comprehensive code search using publicly available crosswalk maps: an ICD-10 to ICD-9 map\(^23\) and an ICD-10 to SNOMED map\(^25\). Our first round of searches identified crosswalk matches for ICD-10’s Z-codes and for all the codes identified using the methods mentioned above. Subsequent rounds used newly identified codes captured from prior searches to find additional crosswalk matches. This was done several times to ensure thoroughness. Code lists obtained from each round of searches were reviewed manually to assess relevancy before inclusion. We also considered alternative mapping if the case could be made that a code had a better fit elsewhere within the taxonomy.

Free-Text Diagnosis Descriptors

While harder to extract, unstructured and character-limited text offers a rich source of SDH data.\(^1\)\(^2\)\(^3\)\(^4\) Documentation of free-text data isn’t necessarily reserved to only the clinical notes section of the EHR. Many EHRs, including PRO and TW, allow free-text capture of diagnoses when a provider is unable to find a code or code descriptor that properly characterizes a patient’s circumstance. In some cases, these free-text diagnosis descriptors may eventually be mapped to discreet medical codes, but most are not.

Table 1 provides an example of how these free-text diagnostic records might appear in a data table within a clinical database for PRO or TW. The descriptor field represents either a code descriptor (when a code is selected by the provider) or a character-limited, free-text diagnosis descriptor (when the provider adds a custom diagnosis).

To capture this free-text data, we text mined the descriptor field within our diagnostic tables using string searches of both keywords and word pairings. Ideas for words and word pairings were obtained from multiple sources: (1) definitions found in published literature\(^12\)\(^14\), (2) common synonyms associated with individual ICD-10 codes\(^5\), and (3) official descriptors for mapped ICD\(^9\) and SNOMED codes\(^2\). Retrieved descriptors were manually reviewed for accuracy. ICD-9, ICD-10, and SNOMED codes that were frequently included in captured records were also collected and evaluated for potential inclusion in the taxonomy. Examples of keywords and word pairing used for text mining have been made available online.
Table 1. Example diagnosis records documents in a TW or PRO EHR clinical database

<table>
<thead>
<tr>
<th>Patientid</th>
<th>icd9</th>
<th>icd10</th>
<th>snomed</th>
<th>descriptor</th>
<th>recorded_dttm</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td></td>
<td>428078001</td>
<td></td>
<td>HYSTERECTOMY; TOTAL</td>
<td>2/12/2013 9:04</td>
</tr>
<tr>
<td>b</td>
<td>574.2</td>
<td>K80.20</td>
<td>266474003</td>
<td>CHOLELITHIASIS</td>
<td>2/22/2014 8:33</td>
</tr>
<tr>
<td>c</td>
<td></td>
<td>171207006</td>
<td></td>
<td>DEPRESSION SCREENING (Date is was done in comments)</td>
<td>2/27/2013 10:55</td>
</tr>
<tr>
<td>d</td>
<td></td>
<td></td>
<td>962</td>
<td>TETANUS</td>
<td>6/6/2013 9:10</td>
</tr>
</tbody>
</table>

Surrogate Data Elements: Health Insurance Status

Clinical data is not the only source of information on a patient’s SRFs; administrative information can also provide a means to indirectly discern this information as well. It is well documented that Medicaid, for example, can serve as a reliable indicator of a patient’s low-income status. \(^{28,29}\) Thus, we explored the use of health insurance data recorded in the EHR to determine if we could use a patient’s coverage status and/or their health insurance carriers to identify patient SRFs.

Results

SDH Framework

We created a framework that is comprised of 97 distinct SRFs organized under 16 domains (see Table 2). Our comparative evaluation showed that the ICD-10 Z-codes were comprehensive, covering a wide scope of SRFs beyond those addressed by the other frameworks. In line with our broad approach, we included all these ICD-10 topics areas even if they were not addressed in other frameworks.
### Table 2. Final SDH taxonomy

<table>
<thead>
<tr>
<th>Domains</th>
<th>Determinants</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ACCESS TO CARE</strong>: Factors that affect a patient’s ability to obtain effective healthcare</td>
<td></td>
</tr>
<tr>
<td>INSUFFICIENT SOCIAL INSURANCE AND WELFARE SUPPORT <strong>†</strong></td>
<td>PROBLEMS RELATED TO TRANSPORTATION *</td>
</tr>
<tr>
<td>PROBLEMS RELATED TO HEALTH LITERACY *</td>
<td>PROBLEMS RELATED TO HEALTHCARE AFFORDABILITY *</td>
</tr>
<tr>
<td>PROBLEM RELATED TO MEDICAL FACILITIES AND OTHER HEALTH CARE ³</td>
<td>UNAVAILABILITY AND INACCESSIBILITY OF HEALTH-CARE FACILITIES *</td>
</tr>
<tr>
<td><strong>Child-rearing</strong>: Factors that affect early child development and the ability to raise children</td>
<td></td>
</tr>
<tr>
<td>CHILD IN WELFARE CUSTODY ⁶</td>
<td>HOSTILITY TOWARDS AND TABOOING OF CHILD ⁵</td>
</tr>
<tr>
<td>INSTITUTIONAL UPBRINGING ⁷</td>
<td>INADEQUATE PARENTAL SUPERVISION AND CONTROL ⁸</td>
</tr>
<tr>
<td>OTHER PROBLEMS RELATED TO UPBRINGING ⁸</td>
<td>INAPPROPRIATE (EXCESSIVE) PARENTAL PRESSURE ⁹</td>
</tr>
<tr>
<td>PROBLEMS RELATED TO MULTIPARTY ⁸</td>
<td>PARENTAL OVERPROTECTION ⁸</td>
</tr>
<tr>
<td>PROBLEMS RELATED TO UNWANTED PREGNANCY ⁸</td>
<td>PARENT-CHILD CONFLICT OR ESTRANGEMENT ⁸</td>
</tr>
<tr>
<td>UPRISING AWAY FROM PARENTS ⁸</td>
<td></td>
</tr>
<tr>
<td><strong>Education</strong>: Factors related to education attainment</td>
<td></td>
</tr>
<tr>
<td>EDUCATIONAL MALADJUSTMENT AND DISCORD WITH TEACHERS AND CLASSMATES ⁸</td>
<td>OTHER PROBLEMS RELATED TO EDUCATION AND LITERACY ³</td>
</tr>
<tr>
<td>FAILED SCHOOL EXAMINATIONS ⁷</td>
<td>SCHOOLING UNAVAILABLE AND UNATTAINABLE ⁷</td>
</tr>
<tr>
<td>LITERACY AND LOW-LEVEL LITERACY ⁷</td>
<td>UNDERACHIEVEMENT IN SCHOOL ⁷</td>
</tr>
<tr>
<td><strong>Employment</strong>: Factors related to employment attainment and work environment</td>
<td></td>
</tr>
<tr>
<td>CHANGE OF JOB ⁸</td>
<td>STRESSFUL WORK SCHEDULE ⁸</td>
</tr>
<tr>
<td>DISCORD WITH BOSS AND WORKMATES ⁸</td>
<td>THREAT OF JOB LOSS ⁸</td>
</tr>
<tr>
<td>OTHER PROBLEMS RELATED TO WORK ⁸</td>
<td>UNCONGENIAL WORK ENVIRONMENT ⁷</td>
</tr>
<tr>
<td>OTHER PROBLEMS RELATED TO EMPLOYMENT ⁷</td>
<td>UNEMPLOYMENT, UNSPECIFIED ⁸</td>
</tr>
<tr>
<td>SEXUAL HARASSMENT ON THE JOB ⁸</td>
<td></td>
</tr>
<tr>
<td><strong>Finance</strong>: Factors related to income sufficiency</td>
<td></td>
</tr>
<tr>
<td>EXTREME POVERTY ⁸</td>
<td>LOW INCOME ⁸</td>
</tr>
<tr>
<td>LACK OF ADEQUATE FOOD AND SAFE DRINKING WATER ⁸</td>
<td>OTHER PROBLEMS RELATED TO ECONOMIC CIRCUMSTANCES ⁸</td>
</tr>
<tr>
<td><strong>Finance/Housing</strong>: Unspecified factors that are related to finance and/or housing</td>
<td></td>
</tr>
<tr>
<td>PROBLEMS RELATED TO HOUSING AND ECONOMIC CIRCUMSTANCES, UNSPECIFIED ⁸</td>
<td></td>
</tr>
<tr>
<td><strong>Housing</strong>: Factors related to housing attainment and suitability</td>
<td></td>
</tr>
<tr>
<td>OTHER PROBLEMS RELATED TO HOUSING CIRCUMSTANCES ⁸</td>
<td>INADEQUATE HOUSING ⁸</td>
</tr>
<tr>
<td>HOMELESSNESS ⁸</td>
<td>DISCORD WITH NEIGHBORS, LODGERS AND LANDLORD ⁸</td>
</tr>
<tr>
<td>LIVING IN HIGH RISK LOCATION ⁸</td>
<td></td>
</tr>
<tr>
<td><strong>Legal</strong>: Legal circumstances impacting patient</td>
<td></td>
</tr>
<tr>
<td>CONVICTION IN CIVIL AND CRIMINAL PROCEEDINGS WITHOUT IMPRISONMENT ⁷</td>
<td>PROBLEMS RELATED TO RELEASE FROM PRISON ⁸</td>
</tr>
<tr>
<td>PROBLEMS RELATED TO OTHER LEGAL CIRCUMSTANCES ⁸</td>
<td>IMPRISONMENT AND OTHER INCARCERATION ⁸</td>
</tr>
<tr>
<td><strong>Life-cycle transitions</strong>: Factors related to the aging and transitioning between major life milestones</td>
<td></td>
</tr>
<tr>
<td>PROBLEMS OF ADJUSTMENT TO LIFE-CYCLE TRANSITIONS ⁸</td>
<td>PROBLEMS RELATED TO LIVING IN RESIDENTIAL INSTITUTION ⁸</td>
</tr>
<tr>
<td><strong>Occupational exposure</strong>: Occupational exposure to agents that adversely affect health</td>
<td></td>
</tr>
<tr>
<td>OCCUPATIONAL EXPOSURE TO DUST ⁸</td>
<td>OCCUPATIONAL EXPOSURE TO OTHER RISK FACTORS ⁸</td>
</tr>
<tr>
<td>OCCUPATIONAL EXPOSURE TO ENVIRONMENTAL TOBACCO SMOKE ⁸</td>
<td>OCCUPATIONAL EXPOSURE TO RADIATION ⁸</td>
</tr>
<tr>
<td>OCCUPATIONAL EXPOSURE TO EXTREME TEMPERATURE ⁸</td>
<td>OCCUPATIONAL EXPOSURE TO TOXIC AGENTS IN AGRICULTURE ⁸</td>
</tr>
<tr>
<td>OCCUPATIONAL EXPOSURE TO NOISE ⁸</td>
<td>OCCUPATIONAL EXPOSURE TO TOXIC AGENTS IN OTHER INDUSTRIES ⁸</td>
</tr>
<tr>
<td>OCCUPATIONAL EXPOSURE TO VIBRATION ⁸</td>
<td>OCCUPATIONAL EXPOSURE TO OTHER AIR CONTAMINANTS ⁸</td>
</tr>
<tr>
<td><strong>Primary support group</strong>: Factors related to a patient’s immediate family and friends</td>
<td></td>
</tr>
<tr>
<td>ABSENCE OF FAMILY MEMBER ⁸</td>
<td>OTHER SPECIFIED PROBLEMS RELATED TO PRIMARY SUPPORT GROUP ⁸</td>
</tr>
<tr>
<td>ALCOHOLISM AND/OR DRUG ADDICTION IN FAMILY ⁸</td>
<td>PROBLEMS IN RELATIONSHIP WITH IN-LAWS ⁸</td>
</tr>
<tr>
<td>SIBLING RIVALRY ⁸</td>
<td>DEPENDENT RELATIVE NEEDING CARE AT HOME ⁸</td>
</tr>
<tr>
<td>DISAPPEARANCE OR DEATH OF FAMILY MEMBER ⁸</td>
<td>PROBLEMS IN RELATIONSHIP WITH SPOUSE OR PARTNER ⁸</td>
</tr>
<tr>
<td>DISRUPTION OF FAMILY BY SEPARATION OR DIVORCE ⁸</td>
<td>OTHER STRESSFUL LIFE EVENTS AFFECTING FAMILY AND HOUSEHOLD ⁸</td>
</tr>
<tr>
<td><strong>Psychological trauma</strong>: Exposure to crime, violence, or other traumatic events</td>
<td></td>
</tr>
<tr>
<td>ADULT ABUSE, CONFIRMED OR SUSPECTED ⁷</td>
<td>PERSONAL HISTORY OF ABUSE/NEGLECT IN CHILDHOOD ⁸</td>
</tr>
<tr>
<td>ADULT NEGLECT/MALTREATMENT, CONFIRMED OR SUSPECTED ⁷</td>
<td>PERSONAL HISTORY OF ADULT ABUSE/NEGLECT ⁹</td>
</tr>
<tr>
<td>CHILD ABUSE, CONFIRMED OR SUSPECTED ⁷</td>
<td>PERSONAL HISTORY OF FORCED LABOR OR SEXUAL EXPLOITATION IN CHILDHOOD ⁸</td>
</tr>
<tr>
<td>CHILD NEGLECT/MALTREATMENT, CONFIRMED OR SUSPECTED ⁷</td>
<td>PERSONAL HISTORY OF OTHER PSYCHOLOGICAL TRAUMA ⁸</td>
</tr>
<tr>
<td>EXPOSURE TO DISASTER, WAR AND OTHER HOSTILITIES ⁶</td>
<td>UNSPECIFIED ABUSE/MALTREATMENT/NEGLECT, CONFIRMED OR SUSPECTED ⁸</td>
</tr>
<tr>
<td>INTIMATE PARTNER ABUSE/VIOLENCE, CONFIRMED OR SUSPECTED ⁷</td>
<td>VICTIM OF CRIME AND TERRORISM ⁸</td>
</tr>
<tr>
<td><strong>Psychosocial</strong>: Relationship between social factors and psychological health</td>
<td></td>
</tr>
<tr>
<td>DISCORD WITH COUNSELORS ⁸</td>
<td>OTHER SPECIFIED PROBLEMS RELATED TO PSYCHOSOCIAL CIRCUMSTANCES ⁸</td>
</tr>
<tr>
<td>INADEQUATE SOCIAL SKILLS ⁸</td>
<td>STRESS, NOT ELSEWHERE CLASSIFIED ⁸</td>
</tr>
<tr>
<td>SOCIAL ISOLATION, EXCLUSION, OR REJECTION ⁸</td>
<td></td>
</tr>
<tr>
<td><strong>Social environment</strong>: Factors related to a patient's social environment and/or community</td>
<td></td>
</tr>
<tr>
<td>OTHER PROBLEMS RELATED TO SOCIAL ENVIRONMENT ⁸</td>
<td></td>
</tr>
<tr>
<td>PROBLEMS RELATED TO LIVING ALONE ⁸</td>
<td></td>
</tr>
<tr>
<td><strong>Societal/cultural</strong>: Socio-political status that have been shown to impact health within the U.S.</td>
<td></td>
</tr>
<tr>
<td>ACCLIMATION DIFFICULTY ⁸</td>
<td>DEMOGRAPHIC MINORITY ⁸</td>
</tr>
<tr>
<td>PRIMARY LANGUAGE OTHER THAN ENGLISH ⁸</td>
<td>IMMIGRATION/MIGRATION ⁸</td>
</tr>
<tr>
<td>TARGET OF (PERCEIVED) ADVERSE DISCRIMINATION AND PERSECUTION ⁸</td>
<td>LANGUAGE BARRIERS ⁸</td>
</tr>
<tr>
<td><strong>Military/veteran</strong>: Factors associated with having a retired or active affiliation with the military</td>
<td></td>
</tr>
<tr>
<td>MILITARY DEPLOYMENT STATUS ⁸</td>
<td>PERSONAL HISTORY OF MILITARY SERVICE ⁸</td>
</tr>
<tr>
<td>STRESS ON FAMILY DUE TO RETURN OF FAMILY MEMBER FROM MILITARY DEPLOYMENT ⁸</td>
<td>ABSENCE OF FAMILY MEMBER DUE TO MILITARY DEPLOYMENT ⁸</td>
</tr>
</tbody>
</table>

# ICD-10 255-265 ¥ Healthy People 2020 ¦ NAM ψ Kaiser Family Foundation Ŧ WHO

963
Mapped Data Elements

We identified and mapped 2,329 medical codes across three standardized medical vocabularies, 10,896 free-text diagnosis descriptors, and 25 health insurance keywords. Each of these data elements were mapped to a single social determinant. Table 3 provides an aggregate count of mapped data elements by domain. The crosswalk maps were the largest source for standardized medical codes. Approximately 44% (n=852) of SNOMED codes came from the ICD-10 to SNOMED crosswalk map. The ICD-10 to ICD-9 map yielded 37 ICD-9 codes.

Through our mining of health insurance names, we found searching with simple keywords did an excellent job in extracting target patient records. Using words like “Medicaid” and “Veteran” was a more efficient extraction approach as compared to collecting thousands of individual health plan names. These words alone were not entirely perfect. We needed to add the local names used by states for their Medicaid programs (e.g. TennCare, CalOptima, Peach State Health...). We also discovered that health institutions often document uninsured status, homelessness, and immigrant status within health insurance data tables.

Table 3. Counts of mapped data elements by domain

<table>
<thead>
<tr>
<th>Domain</th>
<th>SNOMEDs</th>
<th>ICD-10s</th>
<th>ICD-9s</th>
<th>Free-Text Descriptors</th>
<th>Insurance Terms</th>
<th>Total Unique Data Elements</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSYCHOLOGICAL TRAUMA</td>
<td>277</td>
<td>220</td>
<td>52</td>
<td>2921</td>
<td>0</td>
<td>2300</td>
</tr>
<tr>
<td>SOCIETAL/CULTURAL</td>
<td>430</td>
<td>2</td>
<td>0</td>
<td>1698</td>
<td>2</td>
<td>2014</td>
</tr>
<tr>
<td>PRIMARY SUPPORT GROUP</td>
<td>246</td>
<td>15</td>
<td>14</td>
<td>1718</td>
<td>0</td>
<td>1955</td>
</tr>
<tr>
<td>CHILD-REARING</td>
<td>95</td>
<td>19</td>
<td>12</td>
<td>993</td>
<td>0</td>
<td>1119</td>
</tr>
<tr>
<td>EMPLOYMENT</td>
<td>109</td>
<td>11</td>
<td>3</td>
<td>628</td>
<td>0</td>
<td>751</td>
</tr>
<tr>
<td>PSYCHOSOCIAL</td>
<td>82</td>
<td>5</td>
<td>3</td>
<td>631</td>
<td>0</td>
<td>722</td>
</tr>
<tr>
<td>HOUSING</td>
<td>162</td>
<td>3</td>
<td>2</td>
<td>934</td>
<td>1</td>
<td>762</td>
</tr>
<tr>
<td>SOCIAL ENVIRONMENT</td>
<td>80</td>
<td>4</td>
<td>2</td>
<td>457</td>
<td>0</td>
<td>573</td>
</tr>
<tr>
<td>ACCESS TO CARE</td>
<td>96</td>
<td>4</td>
<td>2</td>
<td>457</td>
<td>2</td>
<td>561</td>
</tr>
<tr>
<td>EDUCATION</td>
<td>97</td>
<td>7</td>
<td>2</td>
<td>932</td>
<td>0</td>
<td>428</td>
</tr>
<tr>
<td>LEGAL</td>
<td>84</td>
<td>4</td>
<td>1</td>
<td>320</td>
<td>0</td>
<td>469</td>
</tr>
<tr>
<td>FINANCE</td>
<td>79</td>
<td>3</td>
<td>1</td>
<td>292</td>
<td>19</td>
<td>384</td>
</tr>
<tr>
<td>LIFE-CYCLE TRANSITIONS</td>
<td>38</td>
<td>3</td>
<td>1</td>
<td>348</td>
<td>0</td>
<td>380</td>
</tr>
<tr>
<td>OCCUPATIONAL EXPOSURE</td>
<td>44</td>
<td>12</td>
<td>0</td>
<td>193</td>
<td>0</td>
<td>249</td>
</tr>
<tr>
<td>MILITARY/ VETERAN</td>
<td>9</td>
<td>4</td>
<td>4</td>
<td>189</td>
<td>1</td>
<td>267</td>
</tr>
<tr>
<td>FINANCE/HOUSING</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>24</td>
<td>0</td>
<td>29</td>
</tr>
<tr>
<td>TOTALS</td>
<td>1909</td>
<td>319</td>
<td>101</td>
<td>10896</td>
<td>26</td>
<td>13250</td>
</tr>
</tbody>
</table>

Machine Readable Formatting

After experimenting with multiple ways of storing our mapped SDH taxonomy, we ultimately organized the information using a nested Python dictionary. The primary keys of the dictionary are the 97 individual SRFs found in the taxonomy. Under each SRF are secondary keys consisting of the five data element categories we examined to which individual data elements (code or text) are linked as mapped dictionary values. Figure 1 provides a visual example of the dictionary’s structure and organization.
Discussion

The SDH framework shown in Table 2 is a result of our comparative evaluation of ICD-10’s Z-codes and SDH frameworks promulgated by DHHS Healthy People 2020, WHO, NAM, and KFF. Our analysis showed that ICD-10’s Z-codes covered most of the determinants and domains found in the four other SDH frameworks. We did find a few notable domain gaps, specifically access to health care, socio-political status, and exposure to violence. We chose to add the first two domains (and associated SRFs) to our framework because they were extensively covered and recommended by 4 of the 5 reviewed frameworks. We included the last domain (and SRFs) because it was prominently featured within the SIREN compendium under the Safety domain. We did not include all proposed SRFs or SDHs under these three new domains because some lacked clarity on how such risk factors could be captured in the EHR. For example, we chose not to include the SDHs quality of care and provider availability under the new Access to Care domain given the uncertainty on how this information could be measured and documented within a patient’s medical record. Finally, we chose not to include other domains gaps like behavioral/biological factors and gender given that these topics are either typically addressed as medical risk factors (rather than SRFs) or the proposed scope of the domain is so broad that a data extract would have limited usefulness in SDH informatics. The full comparative analysis, made available online, outlines all the domains, SDHs, and SRFs found in the five reviewed frameworks and highlights those that were and were not included in the final taxonomy.

We reorganized and consolidated domains and SRFs to create a more efficient and intuitive framework. Changes we incorporated include: (1) separating out Financing and Housing into two separate domains, (2) consolidating all SRFs: (a) related to violence under a new domain called Psychological Trauma, (b) related to military or veteran issues under a new domain called Military/Veterans, and (c) related to raising children under a new domain called Child Rearing, (3) consolidating ICD-10 “catch-all” codes (e.g. those codes with descriptors starting with Other problems related to or ending in Unspecified) within each domain, and (4) moving 10 SRFs to different or new domains where they thematically aligned with domain definitions. For example, we moved the Acculturation Difficulty and Target of (Perceived) Adverse Discrimination and Persecution SRFs to the newly created Societal/Cultural domain which includes determinants regarding minority status, immigration, and language barriers.

In terms of our data elements mapping activities, not all standardized medical codes were mapped as indicated within the crosswalk map. In some cases, we assigned certain codes to SRFs that were a better fit. We also chose not to use 142 SNOMED codes and 11 ICD-9 codes that came from the crosswalk maps for two general reasons: (1) the code descriptors were so vague that there would be uncertainty about whether it represented an actual SRF; and/or (2) the
codes represented concepts that most health care institutions would not consider an at-risk situation. For example, we chose to forego mapping SNOMED codes for *inadequate play space* and *inadequate exercise space* to the Inadequate Housing SRF as recommended by the crosswalk maps. The supplemental code search strategies discussed in the Methods section yielded mostly duplicative results; however, each did, to a lesser extent, identify relevant SNOMED codes that would have otherwise been missed. Collecting associated codes as part of the search strategy for free-text diagnosis descriptors, for example, led to the discovery of SNOMED’s “Country of Birth” and “Main Spoken Language” code series which were, respectively, mapped to *Immigration/Migration* and *Primary Language Other Than English*.

Our search strategy for free-text diagnoses revealed quite a bit of redundancy amongst retrieved descriptors. Many of the descriptors would be considered parallel structures (e.g. *loss of job, job loss, lost his job, losing job…*), and, thus, the actual count of meaningfully distinct diagnosis descriptors is likely much less than 11,000.

Our choice of machine-readable format for the taxonomy has many advantages. One benefit of a dictionary format is it allows secondary keys to remain empty if there are no data elements to map as values. Further, arranging the taxonomy and its mapped data elements in a modular fashion around a determinant, instead of in multiple lists or tables, improved our ability to curate an unwieldy amount of interconnected information. The dictionary can be saved as a Python file or JSON, making it portable and shareable. Conversion to JSON format also makes the taxonomy code agnostic. The dictionary is also customizable; readers can unzip and extract the entire mapped taxonomy or just the SRFs of interest. To help readers understand what data elements can be found within the taxonomy, we used the Python library *python-docx* to produce a content catalog that lists all codes (with official descriptors) and all text by domain and SRF. Readers can find a copy of this content catalog at the same site where they can download the taxonomy itself.

There are a number of ways readers can set up the taxonomy for use in their data platforms. Figure 2 provides a Python example on how we typically access the information, specifically the ICD-10 codes. We first open the *.p* file with Python’s *pickle* library. Then we extract all the ICD-10 codes and their associated SDH/SRF label into a data table. We also create a list of these codes that we feed as inputs for filtering queries run against the database containing our EHR data. For a query involving ICD-10 codes, we’d typically extract unique patient ids, unique provider ids, dates of documentation, and ICD-10 codes for all records that match a code contained within our ICD-10 list. The extract is processed at a data table and then we use a *LEFT JOIN* to merge our taxonomy ICD-10 table (right table) with our database extract (left table), matching on the ICD-10 field. With this join, the records in our extract of patient ids, provider ids, and dates are each tagged with one of the 97 SDHs/ SRFs found within our taxonomy. At this point, one can conduct various aggregations to yield patient counts, provider counts, rankings list, or frequencies.

There are a number of weaknesses with our study that should be noted. First, our keyword search strategy was derived from our understanding of how social concepts may be characterized in English within the EHR. Thus, it is limited by “what we know we know”. It is likely diagnosis descriptors were missed because we are not aware of all the possible ways social concepts may be articulated in words and abbreviations. Because the EHR data set is multi-state, there could also be regional differences in use of words and abbreviations. We envision a future role for natural language processing (NLP) to help us gain an understanding of what might have been missed. Second, the taxonomy in its current form solely focuses on identification of documented SRFs. It does not, however, validate the accuracy of this documentation nor does it account for the transient state of many SRFs. Additional work will be required by health informatics professionals to determine if the patients identified by using the taxonomy have a valid and current SRF. Lastly, the taxonomy is only focused on clinical assessment and is binary in nature (i.e. presence or absence of an SRF). It does not currently include a means to capture a richer level of detail of a patient’s social circumstance (e.g.
severity, current/past interventions, etc…). These are details that will be needed by those health care providers who are tasked with intervening to improve the patient’s social situation. Currently, such a taxonomy does not exist, but plans should be made to incorporate these data elements once efforts to build one, like the work being done by the Gravity Project, come to fruition.

Available Materials

A copy of the machine-readable SDH taxonomy in a Python dictionary and JSON formats has been made available online as an open-source resource for others in the community to use, add to, and refine. It can be found at https://github.com/Veradigm-Life-Sciences-Research/SDoHTaxonomy. At this site, readers will also find an SDH taxonomy content catalog, a full copy of our comparative analysis of SDH frameworks, Python code examples, and a list of keywords and word pairing that were used for text mining.

Next Steps

Future work will focus on using our EHR data set and the taxonomy to gain an understanding of SDH documentation patterns amongst health care institutions using Allscripts EHRs. We also plan to investigate a way to leverage the free-text diagnosis descriptors we collected to build a corpus to use with our NLP work, especially in regard to conducting analytics and information retrieval on alternative characterizations of social concepts and on the EHR’s unstructured clinical notes. We plan to determine the importance of adding free-text diagnosis descriptors to an SDH data extraction strategy. This is a unique contribution to SDH informatics and we hope to quantify the value it adds. We plan to continue expanding the taxonomy with LOINC codes and zip codes/census tracts. For the latter, geographic data elements could serve as additional surrogates to identify at-risk patients, particularly areas such as food deserts, housing instability, violence, and financial inequities. We plan to evolve the taxonomy to incorporate the work of the Gravity Project as it becomes available. Our taxonomy is risk factor and data extraction focused which intentionally complements a subset of the critical work of the Gravity Project. We view it important to adapt it as national standards for documenting SRFs are set by the this consensus group.

Conclusion

Using a large, multi-institutional EHR data set in conjunction with published literature, public health policy documents, and existing crosswalk SDH maps, we assembled a comprehensive, mapped SDH taxonomy designed from a data extraction point of view. This work is unique in that it mapped standardized medical codes, free-text diagnosis descriptors, and surrogate data elements, allowing informatics professionals to search for SRF documentation in multiple areas within the EHR. This multifaceted approach is crucial to characterizing the full scope of the impact of SDHs on an institution’s patient population. The layout of the taxonomy within a machine-readable format enables end-users to efficiently unzip, modify, maintain, and share its mapped content. Our work has the potential to help health care organizations characterize their at-risk populations, a critical first step in addressing social needs that impact their patient’s health and well-being.

References

17 Torres JM, Lawlor J, Colvin JD, et al. ICD social codes: an underutilized resource for tracking social needs. Medical Care 2017; 55(9): 810-816.
Characterization of Electronic Health Record Documentation Shortcuts: Does the use of dotphrases increase efficiency in the Emergency Department?

Rimma Perotte, PhD¹,², Christina Hajicharalambous, DO, MSEd, MS¹, Gregory Sugalski, MD¹, Joseph P. Underwood, MD, MHCDS¹
¹Hackensack University Medical Center, Hackensack, NJ; ²Columbia University Irving Medical Center, New York, NY

Abstract

The problem of clinical documentation burden is ever-growing. Electronic documentation tools such as “dotphrases” were invented to help with the documentation burden. Despite the ubiquity of these tools, they are understudied. We present work on the usage of dotphrases within the emergency department. We find that dotphrases are most often used by medical scribes, they significantly increase note length, and are completely unstandardized as to their naming conventions, content, and usage. We find that there is inconsistent usage across and within providers and that there is much duplication in the dotphrase content. We also show that dotphrases have no effect on the time to complete and cosign a note. Finally, we demonstrate that even when accounting for patient complexity upon presentation, note authorship, and note length – notes with higher dotphrase usage are billed at higher billing levels.

Introduction

Documentation is a key part of healthcare. Clinical notes account for a crucial part of all clinical documentation and have important implications for clinical, medico-legal, and reimbursement. In an ideal world, clinical notes are able to provide an up-to-date, accurate, thorough, useful, organized, comprehensive, and succinct narrative describing the patient care trajectory. Accurate and useful clinical notes are especially important in the Emergency Department (ED) setting as the ED is the gateway to the hospital for the large majority of patients. However, the burden to create high quality clinical documentation is sizeable. Proper documentation can be onerous and time-consuming.

There have been some innovations and strategies employed in an effort to combat this pervasive issue. Many clinical departments have employed medical scribes, utilized dictation software, or experimented with personalized EHR documentation shortcuts. Most EHR vendors have the ability to allow for these shortcuts. The two main vendors that account for over 50% of the US healthcare market (Cerner and Epic) both have this type of tool. The Epic version of this functionality is called SmartPhrase and the Cerner version is called AutoText. Both have been referred to as “dotphrases” as the method for calling upon these shortcuts is by typing in a “.” Here, we focus on the Epic SmartPhrase tool.

SmartPhrases are defined by Epic as “shortcuts for entering text quickly.” They are small pieces of text that start with a period and upon typing them in the note, they expand into a larger body of text. They are a powerful tool – they aim to enable high quality, structured documentation while reducing documentation burden and maintaining natural language in clinical text. Dotphrases, if used appropriately, can reduce the number of abbreviations in the clinical record, making the text more readable. They can also enable more consistency, fewer misspelling, and fewer grammar mistakes in the clinical notes. They can also alleviate the need for word sense disambiguation because of their ability to perform abbreviation expansion. For example, if a provider would normally chart “aa” in the record, it could mean

969
“African American” or “antacid”; however, if the provider has predefined an “aa” dotphrase then when she types “.aa” it automatically populates the term “antacid”. In addition, dotphrases can aid in secondary data reuse by enabling the extraction of structured data within an unstructured clinical note.

Dotphrases have been in use for over a decade and have resulted in many shared resources like crowd-sourced tools for creating dotphrases (www.dotphrase.org) and books about best dotphrases to use. As users are able to define their own dotphrases or use shared ones, the functionality is very flexible. Dotphrases can be very broad or specific for a certain clinical condition; they can contain other elements pulled from the chart (Figure 1) or simply be a single word of text.

Figure 1: An example of a dotphrase for documentation of transitions of care at change of shift. “@” indicates automatic pulling in elements from the record. “****” indicates a field that must be filled in by the author before the note can be completed. May also contain “{}” which indicates that a clinician must fill this space in with a value from a predefined dropdown menu.

Another area in which EHR charting innovations can provide value is in increasing provider efficiency. While there has been some work demonstrating that physicians perceive that dotphrases allow them to be more efficient and consistent, our work aims to use EHR data to study whether there truly is a link between dotphrase usage and provider efficiency. We briefly explain the reimbursement methodology for the ED to illustrate our provider efficiency measures.

In order to be financially healthy as a department, providers need to provide quality patient care in an efficient manner. Compensation for care is measured using the following equation:

\[
\text{Compensation} = \frac{\text{Relative Value Unit}}{\text{Patient}} \times \frac{\text{Patients}}{\text{Hour}} = \frac{\text{Relative Value Unit}}{\text{Hour}}
\]

where a relative value unit (RVU) has a set dollar value amount based on the Current Procedural Terminology (CPT)® that is coded. Because of this equation, you can increase efficiency by either increasing your RVU (by increasing the CPT level that is coded) or by seeing your patients more efficiently. Emergency department visit-level CPT codes are based on the level of complexity of the patient’s care. Visits are assigned CPT codes based on the documentation of their complexity in the EHR. For ED visits there exist 6 different visit-level (99281-99285,
99291) codes ranging from non-urgent to critical care. Codes that reflect higher complexity provide higher RVU and therefore higher reimbursement to the provider.

In most EDs, the codes are assigned by outside coders who use the clinical documentation provided to assign one of the 6 codes for each visit. There are Centers for Medicare and Medicaid Services billing requirements as to the necessary elements that need to be present in the clinical documentation for each visit level. As expected, the required elements increase substantially as the visit complexity level increases. For low level visits (99281), the required documentation of history, physical exam, and medical decision making is quite sparse. For a high complexity encounter (99285), the requirements of documentation are much greater, requiring at least 10 body systems to be reviewed for the history section alone.

Our work aims to explore the relationship between dotphrase usage and a set of factors that can measure efficiency such as: note length, time to note completion, time to note co-signature, and CPT coding level. We also generally examine and describe the dotphrase usage landscape within an organization and dotphrase content similarity.

Methods

Data Acquisition and Processing

This study was conducted using EHR data from the Hackensack University Medical Center adult ED. The adult ED had over 65,000 patient visits in 2020 and treats all patients aged 22 and older.

We conducted a retrospective analysis of all ED clinical notes written from January 1, 2020 - March 4, 2021. The analysis focused on the “ED Provider Note” note-type which is the standard and most comprehensive note written for every patient visit in the adult ED. The clinical note metadata and visit-level information was extracted from the Epic enterprise data warehouse. The note corpus was filtered to only look at completed notes. As any providers in the entire organization can use an ED Provider Note, and may do so by accident on occasion, we filtered the corpus further to only look at notes that were signed by an advanced practice provider (APP) or attending, and that the note signer had at least 100 notes signed over the time period.

In a modern day emergency department, many different parties contribute to clinical documentation. In a large academic department, the routine workflow is for residents to initiate and complete the majority of clinical documentation. In this scenario, attending physicians always review the residents’ documentation and then cosign those notes making any edits they deem necessary. In instances where the resident presence is not as robust, medical scribes may be employed to shadow physicians and document the care that is being provided. Their notes are also reviewed, edited and cosigned by the attending physician or APP. A third scenario in documentation workflow would be for the attending physician or APP to construct the entirety of a note from scratch. Regardless of the pathway, an attending or APP is ultimately responsible for the care of the patient and the makeup of the final ED Provide Note text.

To capture all of these variations in documentation contributors, we separately extracted author and co-signer information. The set of extracted variables included: note author, note co-signer, time of note creation, time of note signing by the note author, time of note signing by the note co-signer, note length, and dotphrases used for the note.

In addition, for each of the encounters accompanying the clinical notes, the triage acuity level and the visit-level CPT code was extracted. The triage acuity level is measured using an emergency severity index (ESI) level which ranges from 1 to 5. An ESI level is a measure of the resources required for an ED patient based on their initial presenting signs and symptoms. An ESI level 1 would require maximal resources to treat in the department and would be the most medically complex patient. Although an imperfect predictor of a patients’ complexity, on the whole it is expected to correlate with the final CPT visit-level code assignment.

Analyses

In our analysis we describe the dotphrases prevalence and usage patterns across the department. To understand whether dotphrases are used uniformly across all users we examine the joint distribution of dotphrase usage.
frequency and frequency of usage for individual providers. We also report on dotphrase usage patterns across different author types (scribes, attendings, APPs, and residents).

To examine the impact of dotphrase usage on documentation we looked at four different measures: note length, time to note completion, level of billing code, and dotphrase template similarity.

Correlations were completed in two different ways, 1) using raw counts of dotphrases per note, and 2) using a threshold to categorize high vs. low dotphrase usage per note. A threshold of 2 was chosen after examining the distribution of dotphrase usage across notes. The correlation between dotphrase usage and note length, time to note completion by the author, and time to note signature by the signer were all measured with a Pearson correlation coefficient.

To assess whether the usage of dotphrases could lead to a higher CPT code level, a multiple linear regression model was constructed. The independent variable was the number of dotphrases used and the dependent variable was the level of CPT code. The model adjusted for length of the clinical note, ESI level, and whether a scribe, resident, attending, or advanced practice provider wrote the note.

Finally, to more thoroughly understand the issue of dotphrase expansion and proliferation, a deep-dive was performed into a set of dotphrases meant to deal with the same presentation. To examine the actual content of dotphrases, we extracted dotphrase template text for any phrase with the title containing the word “syncope”. Similarity between these templates was calculated by transforming the dotphrases into vectors using tokenization, stemming, and term frequency-inverse document frequency (tf-idf) and assessing their cosine similarity.

All analyses were completed using Python v3.7.

Results

Descriptive Statistics

Our corpus contained 82,429 ED Provider notes written during the 14-month period. 239 unique individuals across four different types of providers wrote these notes: scribes, residents, advanced practice providers, and attendings. We had 110 unique clinicians across two different types (advanced practice providers and attendings) who cosigned these notes.

Out of the total 82K notes, 11% of them did not contain a single dotphrase. Within the other 73K notes there were a total of 190,777 dotphrases used (an average of 2.3 dotphrases per note). There were 1,763 unique dotphrases. Most of the dotphrases contained elements meant for pulling information from the medical record (such as MRN, age, sex, etc.) but 3.4% of the unique dotphrases were exclusively text phrases such as “Labs, imaging and EKG reviewed and interpreted by me”.

The most popular dotphrase used was “SANOW”. This phrase auto-fills the current time when the author is writing the note. It was found in 23% of all notes and used by 45% of all writers.

We identified that there does not exist a standard naming convention. Some dotphrase names were simply numbers “5”, “88”, “99”, etc. This naming is presumably for ease of searching and typing. We also found that many dotphrases were named for the user, either with a full name, such as “ANDREW1” or with initials. In addition, we found many users creating dotphrases with their name and the name of a condition, such as “CHSYNCOPE”.

Aside from unconventional naming, we also identified a large amount of duplication. Within this span of 14 months, we found that 20 different dotphrases whose titles contained the term “syncope” were used, and 85 different dotphrases whose title contained “COVID”. The prevalence of user-specific dotphrases is evidenced by the finding that over half of the 1,763 dotphrases were only ever used by one author and only 38 dotphrases were in at least 1% of the notes in the corpus (Figure 2).
We also found that intra-author variability in dotphrase usage was quite high. The average intra-author variability was 1.0 ranging from 0 to 6.5. Medical scribes were the most variable in terms of the number of dotphrases used per note. In addition to increased variability, medical scribes were found to have written the most number of notes (40%) and consistently used double the number of dotphrases than residents, APPs, or physicians (Figure 3).
Figure 3. A violin plot demonstrating the distribution of dotphrase usage per note, by different author types. The plot is scaled by the number of observations in each bin. The data points represent the distribution of the raw number of dotphrases used in each note.

Impact of Dotphrases on Documentation

The usage of dotphrases was found to be positively correlated with note length (Figure 4).

Figure 4. A box-and-whiskers plot demonstrating the positive correlation between the number of dotphrases used and the length of an ED provider note (pearson $\rho =0.52$).

When examining the department as a whole, there was a statistically significant difference in the time to note completion by the author. Less dotphrase usage actually correlated with faster note completion time. However, when conducting a paired t-test and examining each individual author’s time to note signature with high or low dotphrase usage, the effect disappeared.
In addition, when grouping the notes into high vs. low dotphrase usage, there was significant difference in time to clinician signature and triage level (Table 1).

**Table 1.** Comparison of note statistics for notes where there was low dotphrase usage vs. high dotphrase usage. Variables with an asterisk indicate statistically significant differences using a t-test.

<table>
<thead>
<tr>
<th></th>
<th>ED Provider Notes with 0-1 dotphrases Used</th>
<th>ED Provider Notes with 2+ dotphrases Used</th>
</tr>
</thead>
<tbody>
<tr>
<td>N (%)</td>
<td>37,201 (45%)</td>
<td>45,226 (55%)</td>
</tr>
<tr>
<td>Mean (SD) Hours to Completion by Author*</td>
<td>32.5 (238.9)</td>
<td>36.5 (187.3)</td>
</tr>
<tr>
<td>Mean (SD) Hours to Completion or Signature by Clinician*</td>
<td>60.2 (257.7)</td>
<td>43.7 (192.7)</td>
</tr>
<tr>
<td>Mean (SD) note length*</td>
<td>4,866 characters (2422)</td>
<td>7,486 characters (3073)</td>
</tr>
<tr>
<td>Most frequent author type (%)</td>
<td>Resident (60%)</td>
<td>Scribe (68%)</td>
</tr>
<tr>
<td>Mean (SD) triage acuity*</td>
<td>2.7 (0.74)</td>
<td>2.9 (0.74)</td>
</tr>
</tbody>
</table>

A linear regression model was created to test whether higher dotphrase usage was correlated with higher billing codes. The model adjusted for triage ESI level, note length, and author (resident, advanced practice provider, attending, scribe) type. The model found that high dotphrase usage was significantly correlated (p<0.0001) with higher-level CPT codes. Another model was also trained using raw dotphrase counts rather than a binary classification of high or low dotphrase usage and the results did not differ.

**Deep Dive into Disease-Specific dotphrases**

To examine the level of duplication and similarity across dotphrases, a sample of dotphrases was extracted and analyzed. The text of the templates for the 20 dotphrases with the term “syncope” contained in the title were studied.

A practicing ED physician assessed each of the phrases and categorized them as such: 70% represented text for medical decision making or medico-legal language, 15% represented guidelines for syncope management, 15% represented other types of text such as EKG interpretations and review of symptoms. There was remarkable similarity within the dotphrase types. In fact, two of the three “pathway” dotphrases were identical.
Discussion

Dotphrases are a flexible and powerful tool. However, similarly to clinical decision support and note templates, without appropriate governance structures the number and content of dotphrases can balloon and become unmanageable.\(^9,10\) Individual dotphrases created by users who have left the organization can linger, dotphrases with outdated clinical practices can remain in the system, or dotphrases minimally used can be forgotten. Our analysis showed that over a 14 month period, half of all dotphrases were only ever used by one provider and 19% of all the dotphrases were used in a single note. In addition, providers can create their own dotphrases because they are unsure of what exists in the system thereby leading to a lot of unnecessary duplication. Our disease-specific analysis found multiple identical or nearly identical dotphrases for a single presenting condition.

We also found that high dotphrase utilizers were most commonly medical scribes. This is anticipated as many scribe training programs encourage and educate scribes to use dotphrases. In fact, proper scribe training has demonstrable benefits - it has been shown that scribes may be writing higher quality notes than physicians themselves.\(^8\)

Our analysis focused on identifying whether dotphrases were able to demonstrate efficiency gains. We found that there was no evidence for efficiency with respect to time: there was no difference in time of note completion by the author or time to co-signature by the signer. However, we did find that there was a significant effect on efficiency reflected as relative value units per patient: CPT coding level was significantly higher when more dotphrases were used. This is likely because use of dotphrases allows the provider to more easily meet the requirements for the highest level of billing by bringing in required elements into the note with relatively few keystrokes. It is clear that many parts of the EHR are designed to assist with billing and dotphrases are likely no exception.

In general, dotphrases can be useful in a number of ways. Most notably, they can decrease cognitive load by incorporating clinical decision rules, pathways, risk calculators, and common differential diagnoses for frequently encountered complaints. These added benefits can also serve as prompts for early EHR users or trainees. Similarly, dotphrases that establish frameworks for infrequently performed, specialized physical exams can help providers with recall. They can also be used as an avenue to store patient resources such as proper discharge instructions, expectant...
management of common conditions, and contacts for addiction or mental health resources centers. While dotphrases have potential as clinical, teaching, and research tools, they can have their downfalls if used improperly.

As our work demonstrated, dotphrases significantly increase note length. This increase in text within a note can increase the signal-to-noise ratio with extraneous information that does not impact medical decision making or patient care. Providers weary about litigation may lean towards higher utilization of unnecessary dotphrases to justify evaluation and treatment plans that oftentimes prove useless in the further management of the patient. The need to comb through extraneous information may result in confusion and unsafe patients’ transfer of care between providers. Even worse, and similarly to the problem of copy-and-paste, using dotphrases without careful thought can proliferate incorrect information, materials, or text.

Our work shows that although there is a lot of work to be done surrounding dotphrase standardization and duplication removal, there seems to be a positive and measurable effect of EHR shortcuts on the billing aspects of clinical documentation.

Limitations

A limitation of this work is that the study time frame was during a pandemic, where practices may not be equivalent to a pre-pandemic time period. However, we don’t anticipate that this changes the study findings. In addition, as the study is conducted within a single academic medical center on the Epic EHR, there is risk of limited generalizability. However, since dotphrases are so ubiquitous in all EHRs, we believe that this work presents an important finding for the informatics community as a whole.

We recognize that there is no control for this work but it would be improbable for us to turn off dotphrases for some subset of the note-writers. There may be a future state in which this era of documentation will serve as the historical control. For example, if the requirements for visit-level coding are modified to be time-based instead of documentation-based, it is certain that many aspects of documentation and dotphrase usage would be altered.

We also note that we may not be capturing all dotphrase usage. Many clinicians may use dotphrases as quick reference guides to aid in appropriate care, but then may delete the dotphrase as it wasn’t meant to be used as part of the documentation for the patient. We do not think that this practice is very common and since we are studying documentation, it is reasonable to focus on dotphrases that are actually used and saved in the text of clinical notes.

In general, although we aim to measure the effect of these EHR shortcuts on efficiency, and thereby documentation burden, we concede that documentation burden is exceedingly difficult to measure. We note that especially in the ED, it may be difficult to assess actual time spent on documentation only using note creation and note signature time.

Conclusion

We found that dotphrases are often used in emergency department clinical documentation. We also found that dotphrases are often very individualized and are not uniformly used across the department. Dotphrases are highly correlated with note length: the more dotphrases that are used, the longer the clinical note. However, although others have demonstrated a perception of increased documentation efficiency we found that there was no effect of dotphrases on note completion time. We also discovered that even when accounting for triage acuity level, note length, and author type there is a significant effect of more dotphrase usage leading to higher visit billing levels.
References


Feasibility of Mobile and Sensor Technology for Remote Monitoring in Cancer Care and Prevention

Susan K. Peterson, Ph.D., M.P.H., Karen Basen-Engquist, Ph.D., M.P.H., Wendy Demark-Wahnefried, Ph.D., R.D., Alexander V. Prokhorov, M.D., Ph.D., Eileen H. Shinn, Ph.D., Stephanie L. March, M.S., Beth M. Beadle, M.D, Ph.D., Adam S. Garden, M.D., Emilia Farcas, Ph.D., G. Brandon Gunn, M.D., Clifton D. Fuller, M.D., Ph.D., William H. Morrison, M.D., David I. Rosenthal, M.D., Jack Phan, M.D., Ph.D., Cathy Eng, M.D., Paul M. Cinciripini, Ph.D., Maher A. Karam-Hage, M.D., Maria Camero Garcia, M.S., Kevin Patrick, M.D., M.S.

1 The University of Texas MD Anderson Cancer Center, Houston, Texas, USA; 2 University of Alabama at Birmingham, Birmingham, Alabama, USA; 3 Stanford University Medical Center, Stanford, California, USA; 4 University of California-San Diego, The Qualcomm Institute/Calit2, San Diego, California, USA; 5 Vanderbilt-Ingram Cancer Center, Nashville, Tennessee, USA.

Abstract

Objectives. Remote monitoring (RM) of health-related outcomes may optimize cancer care and prevention outside of clinic settings. CYCORE is a software-based system for collection and analyses of sensor and mobile data. We evaluated CYCORE’s feasibility in studies assessing: (1) physical functioning in colorectal cancer (CRC) patients; (2) swallowing exercise adherence in head and neck cancer (HNC) patients during radiation therapy; and (3) tobacco use in cancer survivors post-tobacco treatment (TTP).

Methods. Participants completed RM: for CRC, blood pressure, activity, GPS; for HNC, video of swallowing exercises; for TTP, expired carbon monoxide. Patient-reported outcomes were assessed daily.

Results. For CRC, HNC and TTP, respectively, 50, 37, and 50 participants achieved 96%, 84%, 96% completion rates. Also, 91-100% rated ease and self-efficacy as highly favorable, 72-100% gave equivalent ratings for overall satisfaction, 72-93% had low/no data privacy concerns.

Conclusion. RM was highly feasible and acceptable for patients across diverse use cases.

Introduction

Accelerating the fight against cancer calls for enhanced information technology that will give clinicians access to patients’ therapeutic response data in real-time, provide those patients with information critical to and tailored for self-care, and allow researchers to more easily identify emerging trends surrounding those processes.[1] This goal may be achieved with the implementation of systems that enable remote, and when warranted, real-time monitoring of patients’ symptoms and other health-related outcomes in the context of cancer prevention, treatment and survivorship. Home- and sensor-based technology has provided a continuous assessment method for patients with heart conditions, diabetes, and asthma. [2-6] More recently, studies have shown that remote monitoring of symptoms during cancer treatment using a web-based questionnaire platform, coupled with intervention and support by care providers, was associated with better quality of life, fewer treatment interruptions, and improved survival in metastatic cancer patients. [7, 8]

The growing accessibility of mobile and sensor technology may offer scalable and cost-effective strategies to develop and test remote monitoring systems with the goal of optimizing cancer care outside of the clinic setting.[9, 10] The need for additional research has been cited, particularly to help define for which groups of patients and under which circumstances remote monitoring with mobile and sensor technology is most appropriate and acceptable and effective; this need may be particularly pronounced in oncology, where sequelae can be long-lasting.[11-16]

Envisioning the ability to use home-based mobile and sensor technology to address a wide array of challenging problems related to cancer prevention and treatment, and aiming to identify real-world problems faced by cancer clinicians and patients, we designed a software-based cyber-infrastructure to enable collection, storage, processing, visualization, analysis, and sharing of cancer patient and survivor data from multiple domains.[17, 18] The CYCORE (CYber-infrastructure for COMparative effectiveness REsearch) system was designed to combine data from user-friendly, patient-accessible platforms—including mobile sensors and smartphones—with web-based
data display interfaces from multiple sources, to help clinicians and researchers decipher important trends in research participants that are needed to facilitate assessment and clinical decision-making.

CYCORE was built as a service-oriented architecture (SOA) to provide loose coupling between its various system services, promoting independent development and reuse of software components. SOAs provide the means to offer, discover, and interact with CYCORE’s capabilities, such as acquiring sensor data, storing and linking data from various sensors, processing data, or exporting data to other software packages used for outcomes assessment. The CYCORE cyber-infrastructure (CI) is the underlying computational and data infrastructure, which achieves coherent system integration out of a variety of distributed components and manages the lifecycle of all resources. Furthermore, CYCORE uses a Rich Services (RS) architectural blueprint, a type of SOA suitable to integrating crosscutting concerns. The RS architecture allows for infrastructure services, such as encryption, authentication, authorization, and auditing to be plugged into the architecture without modifying core system functionality. This feature ensures scalability, so CYCORE can grow without changes to the underlying CI as new needs are identified and new users engage with the system. We also continuously assess maintenance, usability, and reliability, and we integrate new device models as they become available to improve system quality.

The CYCORE user interface mandates role-based access and provides tailored views for each role. For example, the clinician interface provides a list of patients monitored by a given clinician and displays all sensor data and patient-reported outcomes in a format that simplifies decision-making. The researcher, on the other hand, may access the tools for study monitoring, patient enrollment, sensors assignment, and data analyses. CYCORE provides an integrated view of data from various sources and the ability to run analyses that correlate data and trigger alerts if required by clinicians.

**Objective**

Our research objective was to evaluate the feasibility and acceptability of using CYCORE for remote data collection, outside of the clinic setting, with cancer survivors in three studies that represented unique oncology settings and that assessed: (1) physical functioning in colorectal cancer (CRC) patients; (2) adherence to swallowing exercises in head and neck cancer (HNC) patients during radiation therapy; and (3) tobacco use in cancer survivors who completed an evidence-based tobacco treatment program (TTP).

**Materials and Methods**

This research was approved by the Institutional Review Board at the University of Texas MD Anderson Cancer Center (MDACC).

**CYCORE system overview.** The CYCORE system combined physical entities, such as sensors and mobile devices, with an underlying computational and data fabric. CYCORE had several data acquisition strategies to assimilate multiple sensor types. The CYCORE system consisted of a small plug-in computer-server with the role of sensor hub in the participant’s home, and a backend CI that presented raw and analyzed data to researchers or clinicians. In the prototype used for this study, we developed and implemented our own sensor hub called the Home Health Hub, which was a physical device that aggregated sensor data collected by the patient and relayed these to the CI over an Internet connection. We also developed a smartphone application to enable patient self-recording of videos that automatically uploaded into CYCORE; these were used to monitor adherence. We integrated a customized electronic patient-reported outcome (PRO) system developed at MDACC to provide an app for patient questionnaires via an Android smartphone or tablet.

The system supported an interface with a wide range of consumer-grade sensors that wirelessly transmit data to the home-based server. The sensors integrated into CYCORE for this study were blood pressure monitor and weight scale (A&D Medical), accelerometer (AwareTech Action Tracker), heart rate (Zephyr Technology Bio-Harness), GPS location, and CO monitor (PICO). The accelerometer, heart rate monitor, and GPS provided continuous data. For consumer-grade sensors, the manufacturer's procedures for determining the device calibration tolerance and for recalibration were implemented prior to each deployment. For validation, we employed sensors that exposed a digital interface to the measured values. Sensor data were aggregated, encrypted, and transmitted to a central CYCORE server, which notifies a server database of new data events. Using a web interface, research staff registered and assigned devices to patients, monitored data transmission, and handled technology issues that arose. CYCORE also supported the creation and use of clinician web interfaces for patient monitoring of data during the duration of the study.[18]

**Study descriptions.** We evaluated the feasibility and acceptability of the CYCORE system in three distinct studies with unique study populations, described below. Table 1 describes the types of sensor and mobile technology used in each study.
### Table 1: Descriptions and purposes of sensors and devices used for remote monitoring of cancer survivors

#### CRC Patients: Physical Activity Measurement

<table>
<thead>
<tr>
<th>Device</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>Miniature plug-in computer for receipt and transmission of sensor data</td>
<td>Collect, encrypt, and transmit data from the accelerometers, and the BP, HR, and GPS devices to CYCORE</td>
</tr>
<tr>
<td>Accelerometers (2), worn daily on waist</td>
<td>Monitor physical activity</td>
</tr>
<tr>
<td>Wireless blood pressure monitor, with cuff</td>
<td>Assess physiological changes related to physical activity</td>
</tr>
<tr>
<td>Heart rate monitor, worn daily using a chest strap</td>
<td>Assess physiological changes related to physical activity</td>
</tr>
<tr>
<td>Global positioning system (GPS), carried in pocket or purse</td>
<td>Monitor movement outside of the home</td>
</tr>
<tr>
<td>Smartphone with application for collection of patient-reported outcomes</td>
<td>Collect patient-reported outcome data related to physical activity adherence over multiple time points</td>
</tr>
</tbody>
</table>

#### HNC Patients: Adherence to Swallowing Exercises

<table>
<thead>
<tr>
<th>Device</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smartphone with application for self-video capabilities and collection of patient reported outcomes</td>
<td>Capture adherence to swallowing exercises via video; collect treatment-related symptoms and self-report of swallowing exercises.</td>
</tr>
</tbody>
</table>

#### Cancer Survivors/TTP: Adherence to Smoking Cessation

<table>
<thead>
<tr>
<th>Device</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small plug-in computer (Home Health Hub) for receipt and transmission of sensor data</td>
<td>Collect, encrypt, and transmit CO data to CYCORE</td>
</tr>
<tr>
<td>Carbon monoxide monitor, hand-held</td>
<td>Measure expired carbon monoxide three times daily</td>
</tr>
<tr>
<td>Smartphone with application for self-video capabilities and collection of patient reported outcomes</td>
<td>Assess adherence to CO monitoring; capture patient-reported outcomes data related to tobacco use</td>
</tr>
</tbody>
</table>

Physical activity to promote optimal quality of life in patients with advanced colorectal cancer (CRC). Home monitoring and receipt of reminders to remain physically active may help patients with advanced CRC maintain physical functioning and reduce symptoms and side effects during periods of acute treatment. Staying physically active and managing treatment side effects may, and in turn, help patients tolerate treatment with fewer interruptions, and potentially increase both the length and the quality of their survival. Since higher levels of physical activity are associated with reduced cancer incidence and recurrence, [19-21] we evaluated whether CYCORE would be feasible and acceptable for assessing activity and treatment side effects in CRC patients.

Adherence to swallowing exercises in head and neck cancer (HNC) patients receiving radiation treatment. HNC patients undergoing radiation treatment are recommended to follow a rigorous self-care regimen at home during their 6 to 7-week course of therapy. This regimen includes adherence to a standard-of-care [22-24] prescription of range-of-motion swallowing exercises intended to reduce long-term radiation treatment-induced swallowing complications, [25-27] including trismus, aspiration, mucositis, xerostomia, loss of taste, and fibrosis of the skin and soft tissue. [27-33] Because adherence to both the performance of swallowing exercises as well as appropriate technique is a critical goal for HNC patients during treatment, we assessed the feasibility and acceptability of capturing patients’ self-recorded videos of their daily swallowing exercises. Such videos could be valuable tools for speech pathologists to review and counsel patients on proper technique and adherence to their prescribed swallowing exercises.

Remote collection of exhaled carbon monoxide measurements to assess smoking cessation. MDACC offers evidence-based smoking cessation treatment to cancer survivors and their family members through its Tobacco Treatment Program (TTP), as tobacco use is a contraindication for some cancer treatments and increases risks for complications from the disease as well as mortality. Treatment includes non-nicotine-based medications, nicotine replacement therapy, and behavioral counseling. A standard method of determining cessation status is
through in-clinic exhaled carbon monoxide (CO) monitor measurements, which are only clinically relevant if collected within a day or so of the last smoking bout, limiting the test’s usefulness in validating less frequent smoking behavior. We assessed the feasibility and acceptability of remotely collecting exhaled CO measurements 3 times daily from cancer survivors who had completed the TTP, and to video-record those measurements as an indicator of adherence.

**Participant eligibility and recruitment.** Eligible participants for all studies were 18 years of age or older, English proficient, and had a prior cancer diagnosis. The CRC study included CRC cancer survivors who had completed surgery at least 8 weeks prior to study entry, and who may or may not have been receiving chemotherapy. Those recruited for the HNC study were currently undergoing radiation treatment but were excluded if they had a current swallowing disorder unrelated to their cancer diagnosis. TTP recruits had a history of any cancer other than non-melanoma skin cancer and were current or former smokers but were excluded if they had a known active substance use disorder or had undergone major surgery during the previous eight weeks. For the CRC and HNC studies, eligible patients were identified from medical records and were recruited in clinic. For the TTP study, candidates were identified from program completion records and were recruited via a mailed letter from the TTP director.

**Study procedures.** Participants were provided with study-specific devices and were instructed to use them at home for two non-consecutive 5-day periods, separated by 2 weeks of non-usage. Following informed consent, participants were trained on the use of their devices.

CRC study participants were asked to record two blood pressure (BP) readings, one sitting and one standing, upon arising in the morning and prior to bedtime in the evening. Using a smartphone provided by the study, they completed morning or evening assessments using a mobile app that included standard measures regarding exercise frequency, type, self-efficacy, social support and physical functioning, plus daily random assessments regarding severity of symptoms (fatigue, pain, trouble concentrating, and mood). CRC participants wore a heart rate (HR) monitor and accelerometer during waking hours and used a global positioning system (GPS) device. HNC study participants used a study-provided smartphone to video-record all swallowing exercise sessions prescribed by their speech pathologist and, once-a-day, to self-initiate an assessment of symptoms and adherence to swallowing exercises. TTP study participants exhaled into a hand-held carbon monoxide (CO) monitor three times a day and used a smartphone to video-record those breath tests. Prior to each use of the CO monitor, participants also used the phone to self-initiate questionnaires about the number of cigarettes smoked and exposure to second-hand smoke. Additionally, at three randomly-assigned time points, TTP participants responded to phone-prompted questions about fatigue, pain, trouble concentrating, and mood.

Participants’ perceptions of usability, acceptability, and satisfaction were assessed at seven time points across each study via in-person or phone interviews with research staff. After the baseline training, staff administered a four-item measure regarding ease of use and self-efficacy for using each device. On days 2 and 4 of each 5-day device-use period, participants answered a 6-item measure regarding device problems, medical concerns related to the devices, ease of use, ability to use each device (including reasons for not using a device or why using it was difficult), and what was disliked about each device or that reduced the desire to use that device. [32, 33] This measure was re-administered on day 6 following each 5-day device-use period and included additional questions that assessed belief about the usefulness of automatic data provision to their doctor, concern about data privacy, helpfulness of the initial training session and printed instructions, importance of viewing the data collected, confidence in ability to use at home, and overall satisfaction with device use. Open-ended queries, in all but the post-training survey, solicited additional comments about device use.

**Analysis.** We defined indicators of adherence to daily use of devices for each study as the percent of participants who completed the following during 7 of the 10 device-use days: any BP reading (CRC study), any video showing performance of swallowing exercises (HNC study), and any video showing a CO measurement (TTP study). Videos smaller than 1000 KB were not included in those counts as they were deemed to be unusable. The primary outcome was study completion, defined by completion of the final study (week 4, day 6) survey. Other outcomes of interest were usability and acceptability ratings, and adherence to the requirements for daily use of devices. Open-ended responses to survey items completed by patients were analyzed using a constant comparative approach based on grounded theory. [34]

**Results**

Sample size, demographic characteristics and completion rates, which established feasibility, for each study are shown in Table 2. In reviewing study-specific adherence, we found that 47/48 (98%) of CRC participants who completed the CRC study, transmitted BP data on at least 7 of 10 device use days. Of HNC study participants, 30/31 (97%) self-recorded any videos that captured performance of swallowing exercises (mean, 32.7; range, 1-118 of
1000 KB or greater in size) over the 10-day device-use period; while 16/31 (52%) captured videos on at least 7 of those 10 days. Of TTP participants, 48/48 (100%) self-recorded any video of CO monitor usage (mean, 27.7; range, 2-72 of 1000 KB or greater in size); while 43/48 (90%) captured videos on at least 7 of 10 days.

We found a high completion rate for the device usability and acceptability surveys for all three studies (includes missed surveys due to withdrawal): CRC, 316 of 350 possible surveys (92%); HNC, 213 of 259 (82%); and TTP, 328 of 350 (94%). Participants’ mean responses to the final (week 4, day 6) survey about device usability and acceptability are shown in Figure 1. Scores were generally high on ease of use of devices, self-efficacy, and overall satisfaction, and generally low regarding data privacy concerns.

We found commonalities in participant responses—when provided—to the open-ended questions about what they liked most about using the devices at home. CRC study participants most liked being able to see or monitor the BP results (13/45, 29%); being able to use the devices at home (or not having to travel [in general, or to see the doctor]) (9/45, 20%); and that device use was easy or simple (12/45, 27%) and convenient (5/45, 11%). Some example responses from open-ended questions follow: “Using the blood pressure cuff keeps you kind of in tune with what’s happening, blood pressure-wise.” “Someone received the data at the moment versus me having to explain how I felt at the moment. It picked up certain things that I may have forgotten about.” “I think that whenever you’re monitoring that information and you need it, it’s much easier to do it at the house than go to the doctor’s office.”

HNC study responders most liked that device use provided motivation or a reminder to do the swallowing exercises (8/29, 28%); and that using the devices was easy (7/29, 24%) and convenient (7/29, 24%). Answers to the open-ended questions included the following: “It motivated me to do exercises I might have skipped otherwise.” “This is something in the future that the patients can use to keep in touch with their doctors, and their doctors can monitor them at home while they’re going through treatment.” “I did it at my leisure, first thing when I woke up. I was at a hotel.”

TTP responders most liked being able to use the devices at home (or not having to travel [in general, or to see the doctor]) (15/39, 38%); being able to see or monitor the CO results (9/39, 23%); and that it was easy (8/39, 21%) and convenient (6/39, 15%). Some sample statements follow: “If I had to go into an office, it would be tough to get this in three times a day.” “I like the idea that I could see the readings, and I could tell how much how much carbon monoxide was in my body.” “If I could have kept the devices longer, it probably would have worked to wean me off of smoking.”

Table 2. Demographic characteristics and study completion rates for colorectal cancer survivors (CRC), head and neck cancer survivors (HNC) and cancer survivors who completed the Tobacco Treatment Program (TTP)

<table>
<thead>
<tr>
<th></th>
<th>CRC</th>
<th>HNC</th>
<th>TTP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years), M (range)</td>
<td>55 (25-79)</td>
<td>56 (23-78)</td>
<td>54 (29-72)</td>
</tr>
<tr>
<td>Study completion % (n)</td>
<td>96 (48)</td>
<td>84 (31)</td>
<td>96 (48)</td>
</tr>
<tr>
<td>Sex % (n)</td>
<td>Female</td>
<td>50 (25)</td>
<td>27 (10)</td>
</tr>
<tr>
<td>Race/ethnicity % (n)</td>
<td>White</td>
<td>70 (35)</td>
<td>91 (34)</td>
</tr>
<tr>
<td></td>
<td>Black</td>
<td>14 (7)</td>
<td>3 (1)</td>
</tr>
<tr>
<td></td>
<td>Hispanic</td>
<td>8 (4)</td>
<td>3 (1)</td>
</tr>
<tr>
<td></td>
<td>Asian/other</td>
<td>8 (4)</td>
<td>3 (1)</td>
</tr>
<tr>
<td>Education % (n)</td>
<td>&lt; High school graduate</td>
<td>6 (3)</td>
<td>5 (2)</td>
</tr>
<tr>
<td></td>
<td>High school graduate</td>
<td>12 (6)</td>
<td>11 (4)</td>
</tr>
<tr>
<td></td>
<td>Trade/vocational/some college</td>
<td>28 (14)</td>
<td>22 (8)</td>
</tr>
<tr>
<td></td>
<td>College graduate</td>
<td>40 (20)</td>
<td>32 (12)</td>
</tr>
<tr>
<td></td>
<td>Post-college</td>
<td>12 (6)</td>
<td>14 (5)</td>
</tr>
<tr>
<td>Marital status % (n)</td>
<td>Married</td>
<td>68 (34)</td>
<td>97 (36)</td>
</tr>
</tbody>
</table>
Figure 1. Participants’ responses to post-study evaluation questionnaire regarding usability and acceptability of study devices (range, 0= not at all, 10= extremely)

Colorectal cancer survivors (n=48)

0 1 2 3 4 5 6 7 8 9 10

Head and neck cancer survivors (n=30)

0 1 2 3 4 5 6 7 8 9 10

Cancer survivors who completed Tobacco Treatment Program (n=48)

0 1 2 3 4 5 6 7 8 9 10

Legend:
- AT
- BP
- HR
- GPS
- PRO
- HUB
- PRO
- VID
- CO
- Hub
**Legend:**  
AT: ActionTracker, commercial accelerometer; Act: Actigraph accelerometer; BP: blood pressure monitor; HR: heart rate monitor; CO: carbon monoxide monitor; PRO: patient-reported outcomes mobile application; VID: video mobile application; Hub: Home Health Hub

1Post-study evaluation questions included: 1) helpfulness of baseline training session (training); 2) clarity of printed instructions used at home (instructions); 3) ease of device use at home (ease of use); 4) Confidence in ability to use device at home (self-efficacy); 5) usefulness of automatic data provision to doctor (data to provider; N/A for GPS, Hub); 6) importance of seeing device reading at home (viewing data, N/A for Hub); 7) concern about data privacy (data privacy); 8) overall satisfaction with device (overall satisfaction).

**Discussion**

Based on the high study completion rates and participants’ responses to evaluative questionnaires, our findings showed that remote, home-based monitoring for important cancer-related health outcomes was feasible and acceptable in all three groups of patients in our study sample. We found that patients, some of whom were quite ill and were burdened by frequent clinic and hospital visits, willingly used a variety of mobile sensors at home to measure biometric and self-reported outcomes related to physical activity (CRC patients and survivors); swallowing exercise adherence during RT (HNC patients); and smoking cessation adherence (cancer survivors who were former or current smokers). Participants from all three studies were receptive to using the devices and found that doing so was simple, easy, and convenient.

Compared to participants in the CRC and TTP studies, those in the HNC group had the lowest rates of study- and survey completion, and device use. Based on responses to their user questionnaires, a subset of HNC participants indicated that recording the videos took too much time (4/29, 14% of responders), that the recordings were harder to do as treatment progressed and, subsequently, as patients’ treatment burden typically increases, causing them to feel worse (6/29, 21%), and that feedback about content of the videos was nonexistent during the course of the study (3/29, 10%). As these studies were conducted to determine feasibility and acceptability, we did not provide feedback to HNC participants regarding their adherence or quality of performance of swallowing exercises. Future efforts to incorporate rapid, specific feedback to patients that is aimed at helping them improve an important self-care regimen, such as the one recommended for HNC patients, may increase the perceived value, satisfaction and adherence to remote monitoring protocols.[35][36][37]

TTP participants’ responses also supported the value of incorporating clinician review of data and feedback to patients: after the final week of device use, TTP participants indicated that it would be useful for their physicians to be able to monitor the CO and the self-reported data: “It made me more accountable. I smoked less than I usually would because it reported my smoking habits.”

As noted above, the HNC group’s overall satisfaction with using the phone to record videos lessened end of the study; however, their week 4 mean score still indicated that the group experienced high satisfaction with using the phone for that purpose. This is quite impressive, considering that by week 4 of our study, this group was entering week 4 or 5 of RT, a time during which symptom burden is particularly onerous.

While studies like ours can demonstrate the potential feasibility, as well as efficacy, of integrating sensors, mobile devices, and wireless networks with the goal of improving cancer care and patients’ quality of life, concerns about security and privacy related to the use of this technology have been raised. Patients in our study reported relatively low levels of concern about data privacy, which may be an important factor in securing broader acceptability for remote monitoring models such as CYCORE. Patients’ privacy concerns may be minimized if they perceive a value in the use of technology, particularly in partnership with a trusted source, their health care providers. Nonetheless, systems such as CYCORE will likely achieve their potential benefits if the security and integrity of data can be assured.

Our study sample was limited in regard to diversity in race/ethnicity and educational attainment. While our sample reflects the overall patient population at our institution, our findings may not be generalizable to populations experiencing technology disparities. The COVID-19 epidemic illuminated longstanding systemic disparities in technology access and digital literacy.[40] These disparities hindered the ability to access health care during the pandemic through telemedicine, and also impeded early access to COVID-19 vaccines which depended on the ability to navigate mobile apps and internet portals. Persons at risk for experiencing these technological disparities are commonly in rural areas, lower-income neighborhoods and minority communities and often medically underserved and at increased risk for cancer as well as higher morbidity and mortality from the disease; potentially, those who may benefit from remote monitoring to help manage their illness. Future research on remote patient monitoring must focus on understanding how technology access and literacy affects usability and feasibility when implementing such interventions.[41]
Evaluation of CYCORE continues through expansion of its technical capabilities, increasing the number of participants studied, and through additional studies within the cancer care community. Ongoing and completed studies are focused on efficacy of remote monitoring during acute periods of cancer treatment, as well as research to further establish feasibility and define methodology for use in challenging patient populations. [39] These studies are expanding the types of sensor and mobile data collection methods within CYCORE, including widely used commercially available devices such as Fitbits, further demonstrating that the system is agnostic to the type of device used and the health-related application. The goal is to continue to develop a community of users who would benefit from CYCORE’s capabilities for remote monitoring of patient behavior, all in support of optimizing cancer care [38] by objectively assessing treatment adherence, symptoms, side effects, and toxicities. This should lead to greater patient understanding of and, thus, engagement in their own care, and improvements in data collection for clinical trials including capturing data not typically collected that is highly relevant to cancer recovery and survivorship.

**Conclusion**

Our study contributes to a growing body of research on remote patient monitoring in oncology, particularly studies that combine electronic patient-reported outcomes with collection of biometric outcomes using non-invasive digital devices. [42] A system like CYCORE that gathers and integrates patient-generated health information in cancer prevention and treatment research is highly feasible and acceptable to patients, clinician and researchers. It supports the collection of new forms of data on behaviors and symptoms that are needed to fully determine the course of cancer treatment and health outcomes. These successful examples of remote monitoring through sensors in the homes of different group of cancer patients show promise for broadening the scope and quality of data in cancer prevention, treatment and control. They also demonstrate how patients can be enabled to participate actively in their prevention and treatment regimens, an essential component of successful health outcomes.

**References**


Extracting Patient-level Social Determinants of Health into the OMOP Common Data Model

Jimmy Phuong\textsuperscript{1,2}, Elizabeth Zampino\textsuperscript{1,2}, Nicholas Dobbins\textsuperscript{1,2}, Juan Espinoza\textsuperscript{3}, Daniella Meeker\textsuperscript{4}, Heidi Spratt\textsuperscript{5}, Charisse Madlock-Brown\textsuperscript{6}, Nicole G. Weiskopf\textsuperscript{7}, Adam Wilcox\textsuperscript{1}

\textsuperscript{1}Division of Biomedical and Health Informatics, UW Medicine, Seattle, Washington; \textsuperscript{2}University of Washington Medicine Research IT, Seattle, Washington; \textsuperscript{3}Department of Pediatrics, Children’s Hospital Los Angeles, Los Angeles, CA; \textsuperscript{4}Department of Preventive Medicine, University of Southern California, Los Angeles, California; \textsuperscript{5}Preventative Medicine and Population Health, University of Texas Medical Branch, Galveston, Texas; \textsuperscript{6}Dept of Health Informatics and Information Management, University of Tennessee Health Science Center, Memphis, Tennessee; \textsuperscript{7}Department of Medical Informatics and Clinical Epidemiology, OHSU, Portland, Oregon

Abstract

Deficiencies in data sharing capabilities limit Social Determinants of Health (SDoH) analysis as part of COVID-19 research. The National COVID Cohort Collaborative (N3C) is an example of an Electronic Health Record (EHR) database of patients tested for COVID-19 that could benefit from a SDoH elements framework that captures various screening instruments in EHR data warehouse systems. This paper uses the University of Washington Enterprise Data Warehouse (a data contributor to N3C) to demonstrate how SDoH can be represented and managed to be made available within an OMOP common data model. We found that these data varied by type of social determinants data and where it was collected, in the time period that it was collected, and in how it was represented.
Introduction

The COVID-19 pandemic has highlighted the urgency of existing informatics needs around clinical data sharing and the collection and integration of high-quality Social Determinants of Health (SDoH) data. While different organizations (Johns Hopkins, NYTimes) aggregated case data to generate publicly-available reports, data on patients and health care system response was limited. This challenge occurred even though most institutions had these data stored electronically and the potential value of such data sharing occurred at scales higher than it had ever been before. For some, we are still trying to understand their significance and determine the best approach to sharing such knowledge.

To address data sharing needs, the National COVID Cohort Collaborative (N3C) was created as an extension of the National Center for Data to Health (CD2H) project. With funding from the National Institutes of Health (NIH) and support from the National Center for Advancing Translational Sciences (NCATS), as of March 2021, N3C has collected data from electronic health records (EHRs) from 45 health systems across the country on 3.75 million patients, over 903,000 of whom have been diagnosed with COVID-19. Participating organizations provide data on biweekly to monthly schedules, based on extraction rules from the EHRs for COVID cases and matched controls. Each organization transforms data (e.g., patient demographics, visit history, diagnoses, medications, laboratory test results, procedures, vital signs) into 4 common data models -- the Accrual to Clinical Trials (ACT) Network, National Patient-Centered Clinical Research Network (PCORnet), Observational Medical Outcomes Partnership (OMOP), and TriNetX. These data are then submitted to N3C for storage in a secure enclave managed by NIH, harmonized into the OMOP common data model, which can be made accessible for researcher requesting data for analyses related to the COVID-19 pandemic. N3C significantly advances EHR data sharing capability being the first public national data sharing initiative with centralized data extracted directly from EHRs in this way.

Despite the fact that SDoH have been associated with COVID-19 incidence and outcomes, data elements required to study the role of SDoH are thought to be under-collected due to historic inattention to patient SDoH. SDoH are the non-clinical covariates of how people live, grow, learn and age as it relates with how they can manage stressors or prevent worsening health outcomes. As data representations for SDoH are not native to clinical informatics, mapping of detailed information bears a risk of information loss. Patient SDoH may be predominantly captured as diagnosis codes, narrative reports, and structured survey tools. Diagnosis coding, such as Z-codes in ICD-10-CM representing certain assessments, may be more easily modeled and shared across institutions, but will be limited to those conditions where this coding procedure is available and reliably implemented. Narrative text representations may be the easiest to document for clinicians and therefore could have the highest level of completeness and contextual information, but extracting narrative notes may be imperfect and inconsistent for sharing in common data models. Structured forms can be more readily disseminated and queried electronically; however, variations in research instrument used, instrument version, and coverage of SDoH questions and answer options presenting challenges for comparisons across institutions. Each of these approaches are further limited by whether or not SDoH are assessed as part of the standard clinical workflow. Efforts are needed to define how the data are included and how they can be used for N3C research.

Clinical screening tools for SDoH were introduced as research instruments to help care providers identify and address individual-level SDoH and social needs. These structured screening tools can provide a snapshot of non-medical conditions in the patients’ lives, and are broadly associated with patients’ ability to manage stressors and/or avoid worse health outcomes. Reliable collection of SDoH at the point of care can help clinicians to provide personalized recommendations, promoting health and well-being towards the overall population serviced by the health system.

To make SDoH from EHRs usable in research and other analyses, they must be made computable, which requires knowledge of how these data are being collected and represented. Social and environmental determinants of health continue to have a central importance in studying disparities in COVID-19 outcomes and predisposed vulnerabilities. These factors have been used to guide the US vaccine deployment strategy to protect the most vulnerable and at-risk individuals. Information collected from SDoH structured screening tools may be stored within EHRs as FlowSheets measurements, comparable to how depression is screened. FlowSheets offer a tractable tabular data-structure to examine related data-points for analytics purposes. However, FlowSheet measurements are encoded independently. Between health systems, the same measurements (questions/variables) and values (answers provided) will likely contain locally unique encoding and nuances. To facilitate cross-site analyses, sites must first harmonize their FlowSheet information into standard categories while maintaining...
attributions for source data provenance. With few exceptions, encoding of concepts in these forms is not captured in standard terminologies such as LOINC that enable interoperable analyses.

In this study, we characterize how SDoH data are represented in an EHR to assess their use, limitations, and what recommendations are needed to address challenges in using them for research. This was done using data from UW Medicine and University of Washington School of Medicine, which was an early contributor of data to N3C. The intention of this work is to contribute towards advancing data maturity for patient SDoH representation, data utility for research on social needs and health services research, and ongoing N3C SDoH research on COVID-19 efforts.

**Methods**

**Study design**

This case-study specifically focused on secondary use of information from an Epic Clarity Flowsheet structure into an OMOP database schema. We developed a process for triaging FlowSheet measurements for patient-level SDoH then map observational data elements for Extract-Transform-Load (ETL) operations. We focused primarily on SDoH information represented in clinical observations captured in FlowSheet tables, which includes survey responses and patient-provided occupation information. Post hoc analysis aims to measure data density and completeness to identify next steps in data engineering efforts.

**Data source**

Data were extracted from the UW Medicine Health Network Enterprise Data Warehouse (EDW), which are collected from four Hospital/Medical centers (Harborview, Northwest, UWMC-Montlake, Valley) and various outpatient clinics in the Greater Seattle King-County metropolitan area, but patients may be received from various geographic areas in Washington, Wyoming, Alaska, Montana, and Idaho (WWAMI) and neighboring regions. The cohort criteria include patients tested for SARS-CoV-2 RT-PCR viral presence test, tested for SARS-CoV-2 IgG antigen titer assay, and/or received condition diagnosis of COVID-19 between Feb 1, 2020 through Feb 20, 2021. Although UW employees and staff had increased COVID testing protocols, UW employees and staff at the time of testing were excluded from this dataset due to institutional policies protecting employee privacy. These data elements include patient demographics, visit history, diagnoses, medications, laboratory test results, procedures, vital signs and other observations that have been extracted from the medical records into the OMOP common data model.

Motivation to use OMOP schema was informed by N3C and the OHDSI CHARYBDIS design and rationale to create COVID-19 shareable data cohorts for research.

**Assessment for FlowSheet measurement records**

The general workflow can be represented within Figure 1. The lead author (JP) first explored the schema of the Epic Clarity FlowSheet tables to identify necessary entities. Next, the lead author (JP) performed manual review to triage the FlowSheet measurements to identify data elements relevant to SDoH and generate initial concept mappings. We used the UCSF SIREN project social needs categories as reference taxonomic labels for qualitative coding of Epic FlowSheet measurement names and display names. We initiated the Epic FlowSheet measurements coding as inquiries of ‘Housing Insecurity / Instability / Homelessness’, ‘Food insecurity’, ‘Employment’, ‘Education’, ‘Health Care / Medicine Access & Affordability’, and ‘Immigration / Migrant Status / Refugee Status’. Of note, we separated questions/inquiries of ‘Household size’ into its own category and we lumped inquiries related to ‘Annual Household Income’ as a measure under the ‘Employment’ category.
During exploration, the FlowSheets schema should allow for recognizing (A) templates (surveys tools), (B) individual measurements (survey questions), (C) measurement values (provided responses), (D) the visit encounter (instance of data collection), and (E) linkable to the OMOP person instances (subjects reference).

2) Triage the FlowSheet measurements into a mapping table with information about the data type and OMOP observation concept semantic equivalent. Retain the relationship between each measurement, unique measure value, and their equivalent value_as_concept_id in a separate table.

3) Join the tables, apply conversion logic, apply a filter for values occurring more than 10 instances (optional), then insert into the OMOP Observation table.

Harmonization to OMOP

Once it has been established that SDoH information was captured within the FlowSheet structure, we examined the unique measurement values related to each FlowSheet measurement for value data types (e.g., string, numeric) and concept representation. Concept representations were triangulated with the FlowSheet measurement comments, which are clinician memos. FlowSheet measurements with overwhelming singleton values were considered free-text and flagged for separate extraction approaches. Flowsheet measurements and their standardized unique values were searched for within the ATHENA vocabularies repository. Using ATHENA, we identified and viewed OMOP version 5.3.1 vocabulary, which contains standardized concept relationships from biomedical vocabularies such as the Logical Observation Identifiers Names and Codes (LOINC) database and the Systematized Nomenclature of Medicine -- Clinical Terms (SNOMED-CT) vocabulary. We prioritized OMOP standard concepts from the LOINC vocabulary, which have incorporated survey tools with defined ‘has answer’ relationships, which we used to verify the question-answer relationships. OMOP concepts were reviewed and assigned on the basis of the question and answer having adequate semantic representation. For example, ‘3-years’ can be represented with ‘36 months’; ‘how long have you been homeless’ cannot be represented with ‘Current housing status,’ or vice versa. Where LOINC concepts were not available for the question or answer, we broadened the concept representation to the SNOMED-CT and AllOfUs_Columbia vocabulary. The lead author (JP) conducted qualitative member checking with subject matter experts in OMOP data engineering (AW) and concept mapping using LOINC and SNOMED (DM) to ensure accuracy and validity of the mapping interpretations. Mapping discrepancies were reconciled to an agreed upon set of corrections.
The triage results in a mapping table that associates each ‘Flowsheet Measurement ID’ with their OMOP representation of concept ID and source concept ID (Table 1). FlowSheet measurements marked for exclusion would have RUN_SET set to NULL.

<table>
<thead>
<tr>
<th>Flowsheet observation features</th>
<th>Example values</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIREN Social Needs category</td>
<td>Housing Insecurity / Instability / Homelessness</td>
</tr>
<tr>
<td>FLOWSHEET MEASUREMENT ID</td>
<td>1234567890</td>
</tr>
<tr>
<td>MEASUREMENT NAME</td>
<td>UWM R HCHN HOW LONG HOMELESS UD</td>
</tr>
<tr>
<td>DISPLAY NAME</td>
<td>How Long Homeless</td>
</tr>
<tr>
<td>NORMALIZED NAME (optional)</td>
<td>Length of time homeless</td>
</tr>
<tr>
<td>CONCEPT_ID</td>
<td>40482660</td>
</tr>
<tr>
<td>SOURCE_CONCEPT_ID</td>
<td>40482660</td>
</tr>
<tr>
<td>IS_NUMERIC</td>
<td>FALSE</td>
</tr>
<tr>
<td>RUN_SET</td>
<td>SDOH</td>
</tr>
<tr>
<td>Answer options (value_as_concept_id)</td>
<td>● More than 3 Years (45876969)</td>
</tr>
<tr>
<td></td>
<td>● 1 to 3 Years (45876410)</td>
</tr>
<tr>
<td></td>
<td>● Less than 1 year (45883868)</td>
</tr>
</tbody>
</table>

Table 1: Features from the FlowSheet Observations table

Once the mapping table was established, we refined ETL scripts to generate new Observation records. Regular expressions apply logic between 1) FlowSheet measurement IDs that were triaged and 2) the measurement’s unique value sets to decide on the 3) corresponding OMOP value_as_concept_id. For numeric values, like ‘Annual household income’, we retain the raw value in value_as_number then discretize into value ranges, which are encoded as value_as_concept_ids. After the ETL has been implemented, we conducted patient-level random walks to review for accurate concept representations.

Assessment of employment information

During patient encounters, patients may provide or update their occupation titles (e.g., ‘nurse’) employment status (e.g., ‘full time’), which may be separately collected as free-text and time-stamped within patient demographics tables. Occupation and employment circumstances may be expressed in a variety of ways. We referred to the 2018 US Census Bureau Occupation Code list (as of Sept 26, 2019)\(^3\) Level 2 categories to form regular-expression rules for mapping occupations. These Level 2 categories are also represented with OMOP standard concepts, originating from the LOINC National Trauma Data Standard vocabulary. We link these mappings to extract occupation titles to OMOP concepts. For example, with an occupation label of ‘icu nurse’, an observation record would be mapped to OMOP observation_concept_id 36203487 for “Occupation [Type]”, and value_as_concept_id 36308137 (corresponding to “Healthcare practitioners and technical occupations” (3000-3550)”), and the source string would be retained for data provenance in the value_as_string and the observation_source_value. We repeatedly reviewed the occupations and occupation category assignments until saturation was achieved and all interpretable values were accounted for. To prevent uniquely identifying information, we imposed that occupation labels must have a minimum occurrence of at least 10 patients, else the record would be excluded for unacceptable risk.

Results

Out of 4200 FlowSheet measurements reviewed, initial triage detected 35 FlowSheet measurements of relevance to understand patient-level SDoH. Assessment of FlowSheet measurement records resulted in 21 FlowSheet measurements that had interpretable concept representations available in OMOP v5.3.1; however, the 21 FlowSheet measurements were orphan questions without obvious linkage to survey templates, so the source screening tool and version could not be established. 14 FlowSheet measurements had measurement values that were non-interpretable or not amenable for secondary use and, therefore, were removed from further extraction and analysis. These 14 FlowSheet measurements that were removed accounted for less than 1000 observation records. The 21 FlowSheet measurements that were mappable accounted for the gross majority of SDoH Observation records (n=445222).

As of Feb 20, 2021, the UW COVID-19 OMOP Limited dataset includes 133833 patients that have been assessed for COVID-19. Of the patients assessed for COVID, 83535 patients had pre-COVID medical conditions,
observations, vitals, and medical procedure information, the timeframe between Jan 1, 2010 through Dec 31, 2019 (Table 2). In contrast, 124021 patients who were assessed for COVID had at least 1 medical record generated since COVID-19 reached epidemic scales, between Jan 1, 2020 through Feb 19, 2021. The number of patients with SDoH observations ranged from 11.7% of patients (n=9795) generated during the pre-COVID period to 82% of patients (n=101717) generated since COVID-19 reached epidemic scales.

Of the 133833 patients, approximately 34.8 million patient observations were extracted into the UW COVID OMOP limited dataset as of Feb 19, 2021, where 1.2% of records were SDoH observations (n=445222 total; n=231651 pre-COVID; n=213571 since COVID reached epidemic scales). Since Jan 1, 2020, 81.88% of patients (n=101522) assessed for COVID have provided updates to their ‘Employment’ information. Pre-COVID documentation of patient SDoH was sparse though most frequently collecting information on ‘Housing insecurity / Instability / Homelessness’, followed by ‘Employment’, ‘Education’ and ‘Household size’. The earliest indications of FlowSheet adoption for SDoH observations started in Aug 2010 for ‘Housing insecurity / Instability / Homelessness’. FlowSheet documentation for patient ‘Employment’, ‘Education’, ‘Immigration’, and ‘Household size’ began in Aug 2014 and later included ‘Food insecurity’ in Apr 2018. In contrast, since COVID, the Patients-to-SDoH-observations ratio indicates that knowledge about patient SDoH have increased drastically, especially for Employment information; other SDoH observations have been collected at a slower pace and indicate repeated data collection for a small pool of patients.

<table>
<thead>
<tr>
<th>Patients (% of total patients)</th>
<th>Observations (% of total observations)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>UW COVID OMOP limited dataset (totals)</strong></td>
<td><strong>pre-COVID</strong> (Jan 1, 2010 through Dec 31, 2019)</td>
</tr>
<tr>
<td>SDoH observations</td>
<td>9795 (11.7)</td>
</tr>
<tr>
<td>Employment</td>
<td>1878 (2.24)</td>
</tr>
<tr>
<td>Housing insecurity / Instability / Homelessness</td>
<td>5178 (6.19)</td>
</tr>
<tr>
<td>Education</td>
<td>5309 (6.35)</td>
</tr>
<tr>
<td>Household_size</td>
<td>1149 (1.37)</td>
</tr>
<tr>
<td>Immigration / migrant status / Refugee status</td>
<td>1674 (2.00)</td>
</tr>
<tr>
<td>Food insecurity</td>
<td>179 (0.21)</td>
</tr>
</tbody>
</table>

**Table 2**: Data availability for each SDoH

Since COVID, of the 101552 patients who provided Employment information, approximately 101497 patients described their ‘Employment: current occupation status’. Based on the most recent ‘current occupation status’ information, 34% indicated ‘Full-time’ employment, 27% ‘unemployed’, 18% ‘retired’ and 17% ‘undetermined’, ‘Part-time’, ‘self-employed’ or a ‘student’ status. Only 9.1% of these patients who provided Employment information (n=9344) provided their occupation title. We were able to map 90% of the occupation titles: 46% were ‘retired’, ‘disabled’, ‘unemployed’, ‘self-employed’, or a ‘student’ status. Only 44% of the occupation titles (n=3737) were successfully mapped to the 23 National Trauma Data Standard occupation concepts [https://athena.ohdsi.org/search-terms/terms/44786930], leaving 10% of occupations as unmappable. The discrepancy between 9344 patients and 101552 patients suggests that 91% of patients did not provide their employment occupation titles. Figure 2 depicts the collection of SDoH information for patients within this dataset and the shift in recent months. The recent spike in ‘Employment’ information coincides closely with the phased COVID-19 vaccine deployment (Figure 2b). Collection of SDoH information in all categories since the start of the Pandemic has exceeded or reached within one order of magnitude as the amount collected over the past 10 years.
Discussion

In this study, we characterized SDoH data collected in EHRs at a healthcare delivery institution and stored in the electronic data warehouse. This characterization allowed for queries to extract these data into a common data model that was used for sharing data with a national cohort study. Data were primarily extracted from Epic flowsheet data, which represented assessments in ambulatory care of SDoH. We found that these data varied by type of SDoH data and where it was collected, in the time period that it was collected, and in how it was represented.

The data we collected from flowsheet data involved a process of manually reviewing the data elements that were being used, by ranking them according to frequency and then assessing whether they were SDoH-related. Many of the concepts were available as social history concepts, and standard documentation structures exist in many EHRs in similar ways. However, we found many concepts that were not part of this single approach for collecting data and that often represented multiple assessment tools that may have overlap in content. This is important to recognize with research data sets that may not have specifically identified and queried SDoH data. Unlike more common data elements that are consistently recorded and extracted (e.g., diagnosis and laboratory results), SDoH assessments will be incomplete without focused queries.

Even when attempting to be comprehensive in gathering structured SDoH data, many of the data elements were stored on a minority of patient records. Business and operational concerns, such as subjective screening for program eligibility and clinical and standardized, department-specific intake workflows drive data collection. SDoH observations may be stored on less than 10% of patients seen in health-settings for any specific measure in earlier years, which may be explained by either new programs and policies or by migration from paper to electronic processes. Espinoza et al. provide some understanding to how this occurs with their development of a maturity model on SDoH data. Briefly, the proposed model describes five domains of institutional capacity (data collection policies, data collection methods, technology platforms, analytics capabilities, and operational and strategic impact) across seven detailed levels of maturity adapted for each domain. When SDoH measures are only stored for specific populations conditional on services provided or location treated, it can reflect a combinatorial data completeness problem of assessment, data collection and data accessibility on the different measures.

The change in data availability over time demonstrates these data collection issues and how they reflect the potential use of those data. Employment information as SDoH data is generally pulled from intake forms and demographic information. Until the COVID-19 pandemic, it had been collected for approximately one in 40 patients. During the pandemic, it increased significantly, possibly due to interest in how employment as “essential workers” may affect individuals’ infection risk, employment-linked vaccine eligibility screening, or perhaps a wider concern about employment loss during the accompanying economic effects. There was then a large spike where the majority of patients had employment information documented, coinciding with vaccine deployment. This demonstrated how a targeted need and workflow for collecting data could dramatically change completeness. This was not surprising, but
the magnitude of the change was impressive. This variation also highlights the ways in which standardization into a common data model may lose contextual information related to operational drivers of data collection that may need to be recovered with metadata related to organization, policy, physician practice patterns, and care setting.

Collecting, harmonizing, and tracking SDoH will likely prove incredibly beneficial as addressing SDOH is now promoted to improve population health outcomes and cost savings. An increased understanding of SDOH helps providers connect patients with relevant social services and target vulnerable populations with health-improving social policies and programs. Lofters et al. found that using self-reported SDoH allows primary care centers to identify cancer disparities. Another study screened children for SDoH during well-care checkups and identified that 25% had unmet needs and provided relevant services. Garg et al. conducted a compelling randomized controlled trial that systematically screened all patients using a validated SDoH instrument and saw an increase in the use of community services within their patient population. The work presented in this article provides the basis for ongoing procedure development to address patient SDoH data collection and use for continued benefit.

There remain important limitations to this work. First, this study was done at a single institution in a single region of the country studying SDoH during a period of time where there was increased focus on SDoH. UW Medicine employees and staff were excluded as their COVID testing and prior medical information were considered protected from research use. Data likely looks different in different organizations, though how they differ may reflect more the maturity of SDoH initiatives at the organizations than regional variables. Methods for automated text analysis and taxonomy development may streamline future efforts by analyzing not only text content but incorporating priors from structured and contextual information from FlowSheets. We did not quantify the inter-annotation agreement as a measure for validity; instead, we checked the concept mappings with subject matter experts in EHR data engineering and SDoH data collection. In addition, we studied only data that were collected in structured forms and did not evaluate data completeness compared to information in narrative text reports. The FlowSheet measurements that were mapped did not have survey or version documentation available. Concept mappings and concept coverage are limited to the vocabularies available within OMOP v5.3.1. Other researchers have demonstrated how SDoH variables can be extracted from narrative text and that it significantly increases data completeness. Knowing the extent of the increase could be helpful in determining strategies for increasing SDoH data. Finally, it is unknown whether changes in SDoH data collection represent only changes related to the COVID pandemic and their sustainability. Some monitoring of these metrics will be useful to identify how the data are changing long-term.

Conclusion

Reliable SDoH data are vital for understanding COVID-19 risks and outcomes, as well as for prioritizing medical resources. While there are large amounts of SDoH data available within most EHRs, these data do not conform to common data models, making them difficult to analyze at large scales. There are a multitude of methods for documenting SDoH data. Structured forms in EHR systems can replicate standard assessment instruments, but these forms can vary within and across institutions and may change over time. In order to enable semantic interoperability, we developed a workflow for triaging EHRs for patient-level SDoH then map observational data elements into the OMOP common data model. We found significant data completeness issues, though we identified increases in the collection of specific SDoH elements. Our work demonstrates the feasibility in making SDoH data elements readily available in OMOP data warehouses.

Acknowledgements

This work was partially funded by the National Center for Data to Health (CD2H) grant [NIH/NCATS U24TR002306], supplemental funding from the National COVID Cohort Collaborative [NIH/NCATS U24TR002306-04S3], and grant funding by the Bill and Melinda Gates Foundation [BMGF INV-016910, “COVID-19: Data Analytics on Cases in the Pacific Northwest”].

References


7. Moscrop A, Ziebland S, Bloch G, Iraola JR. If social determinants of health are so important, shouldn’t we ask patients about them? BMJ [Internet]. 2020 [cited 2021 Feb 18];m4150. Available at: https://www.bmj.com/lookup/doi/10.1136/bmj.m4150


28. Gottlieb L. Uses and Misuses of Patient- and Neighborhood-level Social Determinants of Health Data. The Permanente Journal [Internet]. 2018 [cited 2020 Dec 17]; Available at: https://journals.lww.com/10.1097/MLR.0000000000001418


Cognitive Function Characterization Using Electronic Health Records Notes

Adrienne Pichon, MPH1#, Betina Idnay, RN2,3,4#, Karen Marder, MD, MPH3,4, Rebecca Schnall, PhD, MPH, RN2, Chunhua Weng, PhD1
1 Department of Biomedical Informatics, 2 School of Nursing, 3 Department of Neurology, 4Taub Institute for Research on Alzheimer’s Disease and the Aging Brain, Columbia University, New York, New York, USA

Abstract
Cognitive impairment is a defining feature of neurological disorders such as Alzheimer’s disease (AD), one of the leading causes of disability and mortality in the elderly population. Assessing cognitive impairment is important for diagnostic, clinical management, and research purposes. The Folstein Mini-Mental State Examination (MMSE) is the most common screening measure of cognitive function, yet this score is not consistently available in the electronic health records. We conducted a pilot study to extract frequently used concepts characterizing cognitive function from the clinical notes of AD patients in an Aging and Dementia clinical practice. Then we developed a model to infer the severity of cognitive impairment and created a subspecialized taxonomy for concepts associated with MMSE scores. We evaluated the taxonomy and the severity prediction model and presented example use cases of this model.

Introduction
There are an estimated 5.8 million individuals in the United States (US) age 65 and older living with Alzheimer’s disease (AD), with a projected increase to 13.8 million by 20501. As the sixth-leading cause of death in the US, AD is one of the most significant unmet medical needs of our time2. The Food and Drug Administration (FDA) recently approved a disease-modifying treatment, aducanumab, based on the expected drug’s effect on the surrogate endpoint – 18 years since the last FDA-approved treatment3. Clinical trials are the gold standard for providing evidence on the potential harms and benefits of an investigational treatment, but they are time-consuming and expensive. On average, the development of a disease-modifying treatment for AD requires 13 years and costs $5.7 billion4. This highlights how successful AD clinical trials are crucial.

Eligibility prescreening of potential participants is a major bottleneck to successful AD clinical trial recruitment, even though prescreening has resulted in decreased costs incurred by screen failures due to ineligibility and has helped in strategizing recruitment efforts5. One of the challenges is the determination of the level of cognitive impairment, which is a critical component for determining a potential participant’s eligibility to participate in an AD clinical trial5. The 30-item Folstein Mini-Mental State Examination (MMSE) is the most common measure of cognitive function used in AD clinical trials to define the severity of dementia6. Any score of 24 or more (out of 30) indicates normal cognition. Below this, scores can indicate mild dementia (19–23 points), moderate dementia (10–18 points), or severe dementia (≤9 points)7. However, recent MMSE score (i.e., within one year) is not always readily available in the electronic health record (EHR) and when documented is only found in unstructured clinical notes, rendering it difficult to determine without manual inspection8,9. Given the complexity and the long range of trajectory of changes of AD pathological process over time, determining the patient’s potential eligibility to a clinical trial is challenging without a recent MMSE score. This warrants the research team to consider other documented cognitive symptoms (e.g., increased forgetfulness, worsening word finding difficulty) in lieu of a recent MMSE score, which can result in an inaccurate representation of the patient’s level of cognitive impairment due to research staff’s subjective interpretation of the clinical notes10. Hence, a more efficient way to characterize cognitive status for AD clinical trials is needed.

Unsupervised learning approaches and automated search algorithms have been developed to identify clinical subtypes of AD using EHR narrative11-14. Subspecialized terminology based on keywords and phrases from narrative text were also constructed to classify cognitive impairment11. There are also ontologies available to formally represent AD such as the Alzheimer’s Disease Ontology15, the AD Map Ontology16, and the AlzFuzzyOnto17. The Common Alzheimer’s Disease Research Ontology was developed for AD research working to enable integration and comparative analysis of AD research18. The Semantic Web Application in Neuromedicine was developed to build applications for bench scientists initially for, but not limited to, AD research19. However, to the best of our knowledge, there has been no mapping from cognitive function concepts to MMSE score. Our goal is to develop a subspecialized taxonomy for cognitive status characterization and a model for MMSE prediction using related concepts. In this paper, we identified the concepts pertinent to cognitive function measurement in relation to MMSE from clinical notes and mapped them to the Unified Medical Language System (UMLS). We then developed a model to infer the severity of cognitive impairment and used this to construct a novel taxonomy. This paper reports on a computational method for

# Contributed equally, co-first authors

999
phenotyping cognitive impairment using EHR narratives that has the potential to support assorted downstream tasks, such as identifying patients for recruitment of prospective clinical trials, constructing and describing cohorts for retrospective observational studies, and implementation of supportive tools for providers embedded in EHRs and clinical workflows. These advancements could facilitate a personalized approach to care (therapies and supportive services according to the progression of disease) and provide insight into underlying mechanisms of disease to advance precision medicine. This method could be applied in other contexts with clinically relevant score-based proxies.

**Methods**

Data source and sample selection

This study was approved by the Columbia University Irving Medical Center (CUIMC) Institutional Review Board (#AAAD1873). We purposely selected 150 clinical visit notes, each containing an MMSE score, from 118 distinct patients diagnosed with prodromal (amnestic mild cognitive impairment (aMCI)) or probable AD and seen by CUIMC Aging and Dementia clinicians between February 1, 2020 and November 15, 2020. We extracted the following information from their EHR data: diagnosis (aMCI or AD); the date of visit; type of visit (initial or follow up; in-person or telehealth); MMSE during the visit; language used to administer the MMSE; chief complaint; history of present illness (for initial visits); interval history (for follow up visits); neuropsychiatric symptoms; functional abilities assessment; impression; and plans. Initial visit note is included only if a follow up visit note is available that indicates a diagnosis of AD or aMCI. MMSE score (0-30) was used as the label for this analysis. The eligibility of each patient for each visit was determined solely based on the MMSE score for three AD clinical trial protocols, representing phase 1 (NCT03822208), phase 2 (NCT03282916), and phase 3 (NCT03887455) studies respectively and covering a broad range of MMSE inclusion criterion thresholds (i.e., 16-28, 18-30, and 22-30). The workflow is outlined in Figure 1.

![Figure 1. Workflow overview.](image)

Data preparation and concept extraction

Python 3.6 was used for data processing and analysis. R 4.0.2 was used for descriptive statistical computations on demographic characteristics and the heatmap (using ggplot2). The corpus of extracted clinical text was imported and light text preprocessing was performed. A local version of MetaMap along with pymetamap was used to parse the clinical text and extract concepts. The objects that are extracted with pymetamap contain information about the terms tagged, including the UMLS Concept Unique Identifier (cui), semantic types (semtypes), the phrase that triggered the mapping (trigger), negation (1/0 last digit of trigger), scoring information on the quality of the concept match (score), and location codes (location, pos_info, tree_codes). An example MetaMap Concept Object is shown below:

```python
ConceptMMI(index='5', mm='MMI', score=5.18, preferred_name='Ability to Drive', cui='C4050139', semtypes=['[inpr]', triggers=['Driving'-tx-3-"driving"-verb-0'], location='TX', pos_info='137/7', tree_codes='')
```

Feature selection

After concepts were extracted, we experimented with different methods of selecting concepts for inclusion in subsequent modeling. Filtering based on semantic types was a particularly fruitful approach, specifically for excluding irrelevant or off-topic categories and ultimately selecting the ones used in the final model. Clinical expertise was also applied to review the remaining corpus and manually remove any CUIs that were not relevant (e.g., “wellplate”). We decided not to filter based on MetaMap score because even low scores (within relevant semantic types) matched well when manually compared in the patient chart. Finally, concepts occurring in >98% of notes or <2% of notes were omitted for their low salience or low prevalence.

Feature vector

MMSE score (0-30) was used as the outcome variable (label) for the regression model, and each concept feature (extracted from the EHR via MetaMap) was represented as 1 if it was present in the clinical note, 0 if absent, and -1...
if present but negated. Several sociodemographic features were also included in the feature vector, specifically: age (standardized), language (Spanish), and sex (male).

**Model training and evaluation**

**Split dataset.** The full dataset was split 80/20 into a training and test dataset. The training dataset (n = 120) was used to train the model, and the testing dataset (n = 30) was held out and after training was complete, the model was evaluated on this unseen testing set.

**Parameter tuning.** Regression with Lasso is useful for feature selection and modeling at the same time, producing a sparse matrix of features that are relevant in predicting an outcome of interest. The Lasso penalizes each additional model parameter, driving the coefficients towards either 0 or inclusion in the model. In training the model, we experimented with different alpha parameters between 0.05 and 0.4. These parameters resulted in inclusion of more or fewer concept features in the final model, with varying performance on metrics across the data. In the end, an alpha parameter of 0.25 resulted in a reasonable model that was sparse and not overfit or underfit.

**Model parameters and evaluation.** The magnitude and direction of association for final model parameters are returned from the model training and can be used to calculate predicted MMSE scores for a set of feature vectors. The concepts selected by the model were inspected to determine broad categories for important terms, and to examine positive or negative associations between concepts and cognitive status and the magnitude of this association. The R-squared value and Root Mean Square Error (RMSE) were used to evaluate the model, comparing performance on both the training and held-out test data. Sanity checks were performed across a subset of the clinical notes, to verify if concepts extracted and included as important were indeed present in the clinical text and used in the way assumed from the mapping.

**Taxonomy development and evaluation.** The final concepts selected by the model were then used to generate the novel taxonomy by card sorting iteratively among authors (BI and AP) until consensus was achieved, led by BI who has domain experience in this clinical setting. The face validation of the final taxonomy was evaluated by two independent clinicians, a nurse practitioner who sees individuals with AD in clinical and research settings and a lead clinical research coordinator of an AD research center.

**Results**

**Study sample and MMSE-based eligibility**

The study sample (Table 1) included 118 distinct patients corresponding to 150 clinical visits notes (63% female, 50% White, and 53% Non-Hispanic) with mean (SD) age of 74.3 (8.3) years. Of these clinical notes, 49 (33%) visits were conducted in person and 101 (67%) visits were completed via telehealth. A majority of the clinical notes were from follow-up visits (73%). The sample includes a total of 117 (78%) notes that indicate a diagnosis of AD, while 33 (22%) indicate aMCI. Mean (SD) MMSE score was 20.2 (7.1), representing the full range of possible scores. A majority of the MMSE tests were administered in English (82%). Eligible patients based on MMSE are 112 (75%), 106 (71%) and 76 (51%) respectively for the phases 1-3 trials in order. The sample includes 118 clinical visit notes of patients who were deemed eligible to participate in one, two, or all of the research protocols during that particular visit, showing many patients are eligibility for more than one study. Of these notes, 61% (n=71) indicates that the patient was eligible in all the three studies. A total of 32 clinical visits notes document visits with patients who are too cognitively impaired (i.e., MMSE < 16) to be considered for any of the trials.
Final model and concepts selection

A total of 5569 total unique concepts were extracted from the clinical notes for the initial corpus. After expert inspection by BI, a total of 18 UMLS semantic types were included to identify and select relevant concepts (Table 2). After filtering, 1775 unique concepts remained in the corpus. Table 3 shows concepts that were mentioned more than 90 times. The final model (Table 4) includes 40 features, including 39 concepts and age (standardized), with an alpha parameter of 0.25.

Table 2. UMLS Semantic types used to filter relevant concepts.

<table>
<thead>
<tr>
<th>Event</th>
<th>Language</th>
<th>Physiologic Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Family Group</td>
<td>Mental Process</td>
<td>Self-help or Relief Organization</td>
</tr>
<tr>
<td>Finding</td>
<td>Mental or Behavioral Dysfunction</td>
<td>Social Behavior</td>
</tr>
<tr>
<td>Functional Concept</td>
<td>Pathologic Function</td>
<td>Sign or Symptom</td>
</tr>
<tr>
<td>Health Care Activity</td>
<td>Physical Object</td>
<td>Spatial Concept</td>
</tr>
<tr>
<td>Individual Behavior</td>
<td>Phenomenon or Process</td>
<td>Therapeutic or Preventive Procedure</td>
</tr>
</tbody>
</table>

The final input feature vector included 652 concepts and 3 sociodemographic features. There was an average of 108 concepts per record, the minimum was 21 concepts, and maximum was 335 concepts (Figure 2). We extracted a number of concepts from the notes that represents the full range of MMSE score. Figure 3 demonstrates that having a higher or lower MMSE score is not associated with having significantly more or fewer concepts extracted from the note, which may further indicate the existence of different weights among the concepts. Figure 4 shows the distribution of filtered concepts across the 150 notes. Filtered concepts appear an average of 20 times and at most appeared 145 times.

Table 3. Most frequently used UMLS concepts extracted after filtering irrelevant concepts via semantic types.

<table>
<thead>
<tr>
<th>CUI</th>
<th>Preferred Name</th>
<th>Frequency</th>
<th>CUI</th>
<th>Preferred Name</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1518422</td>
<td>Negation</td>
<td>340</td>
<td>C1527305</td>
<td>Feelings</td>
<td>126</td>
</tr>
<tr>
<td>C2584313</td>
<td>Discussion (communication)</td>
<td>210</td>
<td>C0085639</td>
<td>Falls</td>
<td>125</td>
</tr>
<tr>
<td>C2161222</td>
<td>Evaluation procedure</td>
<td>200</td>
<td>C0700287</td>
<td>Reporting</td>
<td>124</td>
</tr>
<tr>
<td>C1512346</td>
<td>Patient Visit</td>
<td>184</td>
<td>C0015576</td>
<td>Family</td>
<td>115</td>
</tr>
<tr>
<td>C1301732</td>
<td>Planned</td>
<td>184</td>
<td>C0150312</td>
<td>Present</td>
<td>110</td>
</tr>
<tr>
<td>C0392747</td>
<td>Changing</td>
<td>154</td>
<td>C2004062</td>
<td>History of previous events</td>
<td>105</td>
</tr>
<tr>
<td>C0700372</td>
<td>Memory observations</td>
<td>151</td>
<td>C0019665</td>
<td>Historical aspects qualifier</td>
<td>105</td>
</tr>
<tr>
<td>C0025260</td>
<td>Memory</td>
<td>148</td>
<td>C0018524</td>
<td>Hallucinations</td>
<td>104</td>
</tr>
<tr>
<td>C0011011</td>
<td>Daughter</td>
<td>145</td>
<td>C0442519</td>
<td>Home environment</td>
<td>102</td>
</tr>
<tr>
<td>C1515187</td>
<td>Take</td>
<td>142</td>
<td>C0242664</td>
<td>husband</td>
<td>101</td>
</tr>
<tr>
<td>C0242665</td>
<td>wife</td>
<td>142</td>
<td>C4553314</td>
<td>Hallucinations, CTCAE</td>
<td>98</td>
</tr>
<tr>
<td>C0589120</td>
<td>Follow-up status</td>
<td>136</td>
<td>C1299586</td>
<td>Has difficulty doing</td>
<td>98</td>
</tr>
<tr>
<td>C1522577</td>
<td>follow-up</td>
<td>136</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C0332257</td>
<td>Including (qualifier)</td>
<td>132</td>
<td>C1299581</td>
<td>Able (finding)</td>
<td>97</td>
</tr>
<tr>
<td>C0262926</td>
<td>Medical History</td>
<td>130</td>
<td>C0686904</td>
<td>Patient need for (contextual qualifier)</td>
<td>92</td>
</tr>
</tbody>
</table>

The final input feature vector included 652 concepts and 3 sociodemographic features. There was an average of 108 concepts per record, the minimum was 21 concepts, and maximum was 335 concepts (Figure 2). We extracted a number of concepts from the notes that represents the full range of MMSE score. Figure 3 demonstrates that having a higher or lower MMSE score is not associated with having significantly more or fewer concepts extracted from the note, which may further indicate the existence of different weights among the concepts. Figure 4 shows the distribution of filtered concepts across the 150 notes. Filtered concepts appear an average of 20 times and at most appeared 145 times.

Figure 2. Number of concepts per note (n=150)

Figure 3. Count of concepts by MMSE among notes (n = 150)

Figure 4. Count of filtered concepts appear in the notes (n=150)

Exploration with dimensionality reduction

We intended to use the concepts extracted from the regression with Lasso modeling to generate the novel taxonomy through unsupervised clustering methods. Before doing so, we inspected potential patterns that may structurally exist in the data by mapping the data from a high-dimensional space to a two-dimensional space with points colored by
outcomes of interest (severity, study eligibility, provider, and language) using Uniform Manifold Approximation and Projection (UMAP). The dimensionality reduction embedding of the feature vector returned from the Lasso (i.e., the final model parameters) is not useful in discriminating between any of the useful outcomes of interest specific to clinical trial eligibility for a particular protocol (Figure 5). In fact, the plots that show the dimensionality reduction embedding using just the sparse feature vector returned from the model did not show any discernable clustering at all.

Figure 5. UMAP Projection of all feature concepts returned from the Lasso by clinical trial protocol eligibility: (a) phase 1 study (NCT03822208); (b) phase 2 study (NCT03282916); and, (c) phase 3 study (NCT03887455).

To explore further, embeddings were generated and visualized for not only the feature vector returned from the model, but also the vector of all relevant features and the input feature vector from the modeling step (Figure 6). A distinct cluster is noticeable when all features are included, and this is still the case for the embedding using only the features that go into the Lasso regression (i.e., after filtering based on semantic type, the cluster still emerged in the embedding). However, this distinct clustering is gone when only features from the final regression with Lasso model were included, suggesting that our model was able to smooth out this underlying structure (and remove concepts that might confound the results).

Figure 6. UMAP Projection of feature concepts by (a) dementia severity, (b) provider, and (c) language of MMSE administration.

After further investigation, the observed difference probably relates to the difference in provider (Figure 6b). The embedding plots suggest that the blue and the tan providers are outliers; it turns out that these two providers work together in one of the AD centers in the department where they see many of the Spanish-speaking patients (Figure 6c). These differences in concepts extracted by provider were not clear upon manual review of the charts; in fact, the charts seemed on the surface to resemble all of the others. When the two providers (who work together) were removed from the dataset, the natural clustering in the dimensionality reduction disappears. However, we decided to keep the two providers in the final analysis because this better represents real-world practice.
**Evaluation metrics**

The plots below (Figure 7) show the true MMSE score vs the MMSE score predicted by the model, for both training and test data. A perfectly accurate model would show dots along the red reference line y=x. Indeed, at higher MMSE scores, the model is fairly accurate and clustered around this reference line. At lower MMSE scores, the model is much further off with the prediction. Considering the R-squared value, the final model explains 60.1% of the variance in the training data and 31.8% of the variance in the test data.

Figure 8 shows the difference between the predicted and true MMSE score, and we again see that lower true MMSE scores were not predicted very well by the model for either the training or test data, but higher scores were not as far off. If the scores were fully accurate, the dots would line up on the red y=0 reference line.

It is to be expected that for both training and test data, lower scores would be predicted higher and higher scores would be predicted lower when compared to the ground truth. The training data has a Root Mean Square Error (RMSE) of 4.6 and the test data returns an RMSE of 5.0. While still a wide margin, a predicted score that is about 5 points off in either direction would still be helpful in prescreening records to accomplish the clinical task.

![Figure 7. True and predicted MMSE score, with Regression + Lasso](image)

**Final model concepts and novel taxonomy**

We constructed a novel taxonomy (Figure 9) by using iterative card sorting of the concepts identified in our model (Table 4) and evaluated its face validity among clinicians in aging and dementia practice. The taxonomy includes concepts within the domains of the MMSE (i.e., orientation, concentration, working memory, memory recall, language, and visuospatial), and concepts representing domains outside of MMSE that are incredibly important and relevant for determining cognitive status functioning decline in AD such as agitation and home environment. The final classes include memory and cognition, activities of daily living, mood and behavior, medical, and descriptors.

The negative coefficient means that if a concept is present, the MMSE score is lower, and a larger magnitude is associated with a larger drop in MMSE score. In our quality check, the following selected concepts were substantiated by an excerpt from clinical notes: concept “Usually Need Help from Another Person for Eating” from excerpt “would not let him (son) feed her”; concept “Unable to Feed Self” from “discussed feeding tube and palliative care”; concept “Has difficulty doing” from “recently having difficulty with calculation”; and concept “Better than Others” from “he thinks his memory is better than his peers”. Further, the standardized age is associated with lower MMSE, which is expected. There are two concepts in the descriptors category that fall under “Spatial Concept” semantic type (i.e., local and adjacent), two concepts under the “Finding” semantic type (i.e., unable and better than others), one concept under the “Therapeutic or Preventive Procedure” semantic type (i.e., change – procedure), and the rest under the “Functional Concept” semantic type.

Further analysis of the data via visualization revealed that the use of concepts, both present and negated, may vary across the levels of cognitive impairment (Figure 10). Concepts pertaining to activities of daily living (e.g., grooming)
were used more in clinical text of patients with severe cognitive impairment. Interestingly, “restlessness” was used across the level of cognitive impairment but increasingly so as the cognitive impairment progresses. Further, the negated concepts were mostly observed in the clinical texts of patients with no to mild cognitive impairment.

Table 4. Final model (output features and weights) used to generate subspecialized taxonomy for cognitive status.

<table>
<thead>
<tr>
<th>Concept</th>
<th>Co-efficient</th>
<th>UMLS CUI</th>
<th>Times used</th>
<th>Concept</th>
<th>Co-efficient</th>
<th>UMLS CUI</th>
<th>Times used</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hallucinations</td>
<td>-0.185</td>
<td>C0018524</td>
<td>26</td>
<td>Has difficulty doing (qualifier value)</td>
<td>-0.146</td>
<td>C1299586</td>
<td>65</td>
</tr>
<tr>
<td>Usually need help from another person for eating</td>
<td>-2.724</td>
<td>C4318483</td>
<td>19</td>
<td>Progressive metabolic disease</td>
<td>-1.696</td>
<td>C0332287</td>
<td>12</td>
</tr>
<tr>
<td>Home environment</td>
<td>-1.514</td>
<td>C1522577</td>
<td>103</td>
<td>Medical appointment</td>
<td>-1.271</td>
<td>C0596893</td>
<td>21</td>
</tr>
<tr>
<td>Unable to feed self</td>
<td>-0.618</td>
<td>C0566415</td>
<td>28</td>
<td>Unable</td>
<td>0.467</td>
<td>C0457083</td>
<td>91</td>
</tr>
<tr>
<td>Agitation</td>
<td>-0.614</td>
<td>C0085631</td>
<td>17</td>
<td>Assisting (procedure)</td>
<td>0.811</td>
<td>C0034770</td>
<td>17</td>
</tr>
<tr>
<td>Probable diagnosis</td>
<td>-0.572</td>
<td>C0332148</td>
<td>14</td>
<td>Preventive monitoring</td>
<td>1.198</td>
<td>C4522046</td>
<td>18</td>
</tr>
<tr>
<td>Unable</td>
<td>-0.51</td>
<td>C1299582</td>
<td>29</td>
<td>Comprehension</td>
<td>2.149</td>
<td>C0013126</td>
<td>48</td>
</tr>
<tr>
<td>Experimental finding</td>
<td>-0.341</td>
<td>C0162340</td>
<td>18</td>
<td>Experimental finding</td>
<td>2.749</td>
<td>C1270972</td>
<td>41</td>
</tr>
<tr>
<td>-0.268</td>
<td>C1270972</td>
<td>64</td>
<td>Medication</td>
<td>2.749</td>
<td>C1270972</td>
<td>41</td>
<td></td>
</tr>
<tr>
<td>-0.220</td>
<td>C0589120</td>
<td>43</td>
<td>Follow-up status</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 9. Subspecialized taxonomy for characterizing cognitive status of patients with aMCI or AD.

Discussion

The current study demonstrates that narrative text from outpatient clinical visit notes of patients diagnosed with aMCI and AD could be instrumental in indicating the level of cognitive impairment for AD patients. Consistent with the literature, we used a data-driven approach to derive five categories of concepts corresponding to MMSE scores: cognitive, functional, behavioral, medical, and descriptors. These findings provide an important, albeit preliminary, foundation for informing expansions of existing concepts related to a specific proxy (i.e., MMSE score).
Our findings regarding the content and classification of cognitive impairment concepts provide a preliminary understanding of potential challenges for using EHR notes in automated prescreening approaches. First, there are concepts that are frequently used with different meanings in different contexts. For example, the term “local” is used to describe the proximity of clinical care (e.g., local neurologist) or a symptom (e.g., local tenderness). Examining documentation patterns across varying points of care (e.g., follow-up visit) and expanding the analysis of terms to other neighboring terms may provide additional information for the disambiguation of these concepts. For example, the term “feed” may not refer to the patient’s ability to feed oneself but of forgetting to feed a pet, which is still critical in classifying the impairment as memory and cognition or activities of daily living but also different than other uses. Further, the frequency of the terms used varies across the level of cognitive impairment progresses. For example, terms related to the concepts “grooming” and “restlessness” are more frequently used in clinical notes of patients with severe cognitive dementia; conversely, the use of terms related to the concepts of “hallucinations” and “tremor” decreased from normal cognition to severe dementia.

Figure 10. Percent of clinical text with concepts across the level of cognitive impairment, by taxonomy category.

Provided the high prevalence, cost, and mortality associated with AD1—and the urgent goal of the National Alzheimer's Project Act’s (Public Law 111-375) National Plan to Address Alzheimer's Disease to prevent and effectively treat AD by 20252—meeting recruitment goals for AD clinical trials have important scientific, clinical, financial, ethical, and policy implications28. Study findings have the potential to improve recruitment rates for AD clinical research and subsequently accelerate further development of an efficacious disease-modifying treatment for AD. Availability of a specialized taxonomy commonly used in AD clinical care documentation has the potential to bolster eligibility prescreening approaches for clinical trial recruitment. Future work to further develop and refine the model across broader range of patients, providers, and institutions should focus on expanding the model’s capacity to utilize clinical text with symbolic knowledge representation such as building into the model ways to include and infer from relevant broader and narrower concepts. For example, the concept “hallucinations” has “behavioral/psychiatric manifestations” as a parent concept and “sensory manifestations” as a grandparent concept, which are useful to interrogate for future inclusion in the model; on the other hand, “substance withdrawal severity” and “hypocalcemia severity” would not be relevant; finally, “general symptom” may be useful but is very broad. The present study derives potential concepts for inclusion in future ontologies and phenotypes, which can serve as the foundation of developing clinical decision support for clinicians and research teams to identify cohorts for AD clinical research and focus their time and effort into other aspects of research and clinical care such as recruitment and patient education29. Future work should examine the predictive accuracy of the terms used in clinical text and how these maps to UMLS concepts in determining AD clinical trial eligibility across different patient samples and EHRs. It would also be interesting to develop a longitudinal account of disease with the approach to computational phenotyping described in this study.

Limitations

The study has some limitations. The first limitation is the use of retrospective data of patients with two specific diagnoses (aMCI and AD) documented in EHR notes by ten specialists in a single subspecialized clinical practice in a quaternary academic medical center. This design strengthens internal validity, which is important in this early stage.
Moreover, there may be local variation in the terminology used within this single clinical practice, and these findings may not be generalizable to other settings or other EHR systems. As there is no mandate for MMSE testing at any particular visit, each provider perceives and documents different information (particularly across provider type), including the MMSE score; some providers may skew towards new patients, initial visits notes may be written by a clinical fellow or nurse practitioner under the provider’s supervision, time spent with patients may vary widely, and the clinic was newly conducting telehealth visits due to the COVID-19 pandemic. Future research evaluating different data sources and settings is needed to understand whether and how documentation patterns specifying cognitive function vary across services and EHRs. Additionally, while the present study included important sociodemographic factors such as age and language, other relevant factors such as education and date of the previous MMSE were not included due to dispersed EHR phenotypes and fragmented EHR data. As such, future research should investigate these factors using a much larger and more representative corpus including various clinical texts including discharge summaries and clinical notes.

As this study explored concepts in the clinical visit notes of patients diagnosed with aMCI or AD, these concepts may only represent concepts that clinicians who subspecialize in aging and dementia are more accustomed to using. Future research is needed to determine whether the concepts in this taxonomy are consistently used by clinicians across a different range of specialties (e.g., general neurology, primary care) and diagnoses. We could also expand the data corpus to include notes from general medicine and could consider using pretrained embedding models as an alternate approach. The present study focused on identifying terms used by clinicians to describe cognitive symptoms based on MMSE. However, other comorbidities such as stroke, epilepsy, or other neurodegenerative diseases may also affect cognitive symptoms in addition to AD. Future prospective research should aim to identify terminology that may be unique for AD or AD alongside other comorbidities. Further, we did not look at the presence or absence of AD biomarkers result to further confirm the probable diagnosis, and the medications that the patients were taking during the MMSE testing, which may affect the patient’s performance. Finally, as this study represents an initial derivation of a subspecialized taxonomy from a gold-standard diagnostic group only, although the whole MMSE range was covered, it did not include matched controls. Further development and refinement of subspecialized taxonomy to characterize cognitive function will benefit from identification of common data features in the EHRs of patients with subjective cognitive complaints, which was not addressed in this study.

Conclusions

In this study, we introduced a subspecialized taxonomy based on concepts in clinical text to assist in characterizing cognitive impairment without using MMSE scores. Our work demonstrates the feasibility of subgrouping of patients using their EHR notes even when MMSE scores may not be directly available. By leveraging clinical narrative notes, a proxy of cognitive impairment was constructed using symbolic knowledge representation and computational modeling. This method could improve the efficiency and accuracy of cohort identification based on cognitive function for AD clinical research regardless of the presence of a recent MMSE score in the patient chart. We conclude that utilization of specialized taxonomy is a suitable approach to extract concepts from clinical notes and this approach may be more portable and generalizable than a purely computational approach and could possibly be used to infer the severity of cognitive impairment and provide interesting clinical insights. Future work is warranted to test how this approach may generalize to other domains for developing proxies for clinically relevant indicators or formal scores using narrative clinical notes and symbolic methods.

Acknowledgments

Research reported here was supported by the National Library of Medicine grants R01LM009886 (PI: Weng), 5T15LM007079 (PI: Hripcsak), the National Institute of Nursing Research grants T32 NR007969 (PI: Bakken) and K24NR018621 (PI: Schnall). The content is solely the responsibility of the authors and does not represent the official views of the National Institutes of Health. We thank Oliver Bear Don’t Walk IV, Harry Reyes Nieva, and Michael Zietz for helping in debugging and improving our modeling pipeline; Wendy Gonzalez and Arlene Mejia for the face validation of the taxonomy; and Dr. Noémie Elhadad for her insightful comments on the manuscript.

References


NetworkSIR and EnvironmentalSIR: Effective, Open-Source Epidemic Modeling in the Absence of Data

Madison A. Pickering, B.S.1,2, Subbarayan Venkatesan B.S., MS., Ph.D1, Christoph U. Lehmann, M.D.2, Sameh Saleh, M.D.2, Richard J. Medford, M.D.2
1The University of Texas at Dallas, Richardson, Texas; 2University of Texas Southwestern Medical Center, Dallas, Texas

Abstract
The rapidly changing situation characterized by the COVID-19 pandemic highlighted a need for new epidemic modeling strategies. Due to an absence of computationally efficient models robust to paucity of reliable data, we developed NetworkSIR, a model capable of making predictions when only the approximate population density is known. We then extend NetworkSIR to capture the effect of indirect disease spread on the progression of an epidemic (EnvironmentalSIR).

Introduction
The COVID-19 pandemic has led to over two million deaths worldwide.1 While the recent emergency authorization of effective vaccines has signaled a potential end to the spread of SARS-CoV-2, the pandemic has emphasized the importance of understanding the mechanisms that drive epidemic development, to better manage future pandemics. Computer models and/or simulations can provide useful insights on disease spread in a given population such as predicting or modeling the effect of public health policies (such as self-quarantining) without having to enact them2.

While studying disease spread has traditionally been the domain of epidemiologists, public health experts and, more recently, machine learners, there are clear benefits to examining this issue through the lens of distributed systems (a subfield of computer science). The numerous agents within a distributed system can exploit the inherently parallel nature of disease modeling, allowing for massive increases in computational efficiency compared to traditional, serial methods. These advances in efficiency can allow researchers to perform larger and more complex simulations—which consider more complexities—yet require similar resources.

To demonstrate the utility of a distributed systems approach to modeling, we propose and justify a basic implementation of Kermack and McKendrick’s SIR1 (susceptible, infected, removed or recovered) model as a distributed system, dubbed NetworkSIR. We then highlight NetworkSIR’s extensibility by including a modification to create EnvironmentalSIR, which is able to capture a different phenomenon of interest, namely the indirect disease (without direct person to person contact) spread which occurs through objects or air. To further facilitate scientific research on this subject the source code of both models has been made publicly available under the GNU General Public License v3.1

Lastly, while the problem of studying disease spread has persisted for centuries, many of the ideas presented here are new. We believe that NetworkSIR is the first distributed mechanism of simulating disease, and EnvironmentalSIR is the first network-based model capable of capturing indirect disease spread as it is transmitted from the environment to a human3, as opposed to exclusively from human to human. Understanding the effect of environment-to-human based disease spread is particularly important, as it is not currently well understood in the context of computational epidemic modeling. Having a model that can elucidate the effect of such transmission could play a vital role in guiding policy decisions regarding visiting public spaces and restaurants during epidemics.

Previous Work
Kermack and McKendrick developed one of the first and most important population disease transmission models which was both simple and reasonably accurate. The proposed SIR model was compartmental in nature; the population of interest is partitioned into compartments, or categories, based on the characteristics of the population. Specifically, a population was partitioned into those who were ‘Susceptible’ to the disease, those who had become

1 Please refer to https://madisonpickering.github.io/publications_pdfs/NetworkSIR_EnvSIR_repoLinks.txt to access the code.
2 An example of disease spread through the environment to a human is disease spread by touching a surface with infectious material, then touching one’s mouth, nose, or eyes and becoming infected.
‘Infected’, and those who were ‘Removed’ from the population secondary to past infection, where “Removal” included death or immunity to the disease. Epidemics were then “simulated” by using differential equations to show the flow of individuals between compartments.

The SIR model describes the occurrence of an epidemic as depending on two factors:

1. The inherent “infectivity” of the disease
2. The population density of the susceptible population

As many disease-specific factors contributing to its “infectivity” are difficult or impossible to modify by humans, considerable research has been devoted to understanding the effect of non-homogeneous population density on epidemic spread for which it is necessary to understand human social structure. An Infected individual with a broad social network may infect more individuals (simulating the effect of a higher population density) than an individual who may only interact with a small number of contacts.

The tendency of humans to interact with others based on their own personal preferences is referred to as non-homogeneous mixing. Homogenous mixing in contrast considers that a contact between any two individuals occurs randomly, but uniformly, across the population. From a modeling perspective, the effect of non-homogeneous mixing is reduced as a population size approaches infinity, at which point homogenous mixing may be assumed. As we will focus only on the more common scenario of a non-infinite population size, we will similarly confine our discussion of previous work to three broad paradigms which each reject the homogenous mixing assumption.

Firstly, we discuss Cellular Automata based approaches, which build off the concept of Cellular Automata as first pioneered by John Von Neumann and Stanislaw Ulam. In this method, epidemiological simulations take place on some finite, two-dimensional grid, where each point or cell within the grid is represented as a finite state automaton. Each automaton can represent an individual or subpopulation, depending on the way the simulation is described. The automaton use the states (e.g., Susceptible or Infected or Removed) of a small number of their physical neighbors to derive their next state. This approach’s key limitation is its restricted capacity to incorporate non-random mixing. Typically, each individual or subpopulation is allowed to mix with either four or nine neighboring cells: the automaton’s Moore’s Neighborhood or Von Neumann Neighborhood.

In contrast, Bayesian Network approaches (of which we include Hidden Markov Models as a subset) are able to incorporate non-homogeneous mixing without any restrictions. These approaches typically model individuals as nodes in the Bayesian Network and include edges between individuals which represent their capacity to infect another. Only a node that has an edge to another node may directly infect that other node. A number of features (essentially, any recorded data the researcher thinks are important, for example age) are associated with each node, and infection is determined by making probabilistic predictions about the likelihood that a node will infect another. Using a series of these predictions, a Bayesian Network can simulate an epidemic with an impressive degree of specificity.

While Bayesian Network approaches are the most powerful in terms of realistic predictive ability, they also require a large amount of individual-level input to make these predictions. Furthermore, most operations (e.g., training, prediction) with Bayesian Networks are prohibitively computationally expensive as the network grows in size, limiting their current utility to very small scale simulations. Lastly, Bayesian Networks require that no cycles (loops) exist in their node topology. Despite these restrictions, Bayesian Network approaches can still produce impressive results.

Finally, we discuss Contact Network based approaches, which reject homogeneous mixing at a level between Bayesian Networks and Cellular Automata models. Like Bayesian Networks, Contact Networks model humans as nodes and require an edge between two nodes if one individual may potentially infect the other (assuming that one of those humans is infectious). As a result of defining interactions in this manner, Contact Network based models place no restrictions on the amount of non-homogenous mixing that can be represented. This potentially allows for very realistic simulations at an individual level, similar to Bayesian Network approaches. Furthermore, Contact Network based models are not inherently computationally prohibitive, since the asymptotic time complexity of simulating an epidemic is largely up to the researcher. Many researchers have attempted to adapt the SIR model to such a network due to the SIR model’s relative accuracy despite its simplicity. However, there remains little consensus on how exactly SIR should be adapted to a network. We thus propose NetworkSIR as a way to provide a common ground for researchers interested in epidemic modeling using Contact Networks.
NetworkSIR

The general idea behind NetworkSIR is to use the social structure of a population as input, then apply SIR dynamics with respect to that social structure. By social structure, we refer to humans and their daily close-proximity interactions (CPIs). In a divergence from traditional methods of adapting SIR dynamics to a network, we explicitly model the infectious material responsible for infection as small tokens called “agents”. For an Infected individual to infect a Susceptible individual, that Infected individual must pass an agent to the Susceptible individual along its edge. This mode of modeling infection was inspired by the rumor-spreading algorithm models proposed by Giakkoupis et. al. and its verisimilitude with viral or bacterial vectors.

NetworkSIR simulates an epidemic as follows:

1. **NetworkSIR** uses the social structure of a population of interest as input. This social structure is specified as a list of nodes and edges, where each node represents a human and each edge represents a close proximity interaction (CPI) between those two humans.
2. The entire population of interest is initialized as Susceptible, with the exception of a very small number of individuals who are Infected.
3. The simulation then occurs in a series of rounds, where each round is equivalent to some time period (e.g., a day). In each round, the following transitions occur:
   a. The individuals who are Infected have some probability to infect the nodes that they have edges with. This is done by passing agents from the Infected node along the node’s edges. The mechanism behind this is discussed in more detail later.
   b. A node that has been Infected transitions to being Recovered after some number of rounds, denoted as the Recovery Threshold.

The simulation ends when either no further infection is possible or a predetermined round limit has been reached. Figure 1 depicts a small sample simulation.

![Figure 1: Simulation of NetworkSIR With Three Individuals and Recovery Threshold = 2](image)

**Figure 1.** A sample simulation of NetworkSIR

It is impossible for further infection in the system to occur beyond the third round; once a node has been infected, it cannot become reinfelected. Consequently, once node A receives the agent passed to it by node B, it will simply discard it. Furthermore, Node C cannot directly interact with Node B, due to the lack of an edge between them. While it is not shown here, the simulation will actually run for another round until node B recovers for reasons related to computational efficiency. Intuitively speaking, it is less computationally intensive for each node to report if it is Recovered than for that node to perform a search on the nodes around it and determine if it is in some path that could result in further infection. Furthermore, each round only runs for as long as is required computationally. In a round where no infection is possible, the nodes will simply begin the round, nothing will happen, and then they will terminate.

A major shortcoming of Contact Network based models is the lack of consensus regarding the implementation of SIR dynamics within the network. This typically manifests as individual research groups defining their own equations for how to simulate the compartmental transition from Susceptible to Infected (the process of infection)
and the transition from Infected to Recovered. While this is not inherently problematic, the equations used to govern those transitions usually lack robust justification and source code is often not available\textsuperscript{6, 9, 10, 11, 12}. This makes comparison and evaluation among models difficult or impossible.

By providing extensive justification\textsuperscript{2} for the mechanisms that NetworkSIR uses to simulate an epidemic, we hope to inspire a “common ground” for Contact Network based models going forward. Specifically, we provide justification for the two transitions responsible for simulating SIR dynamics in NetworkSIR, based on real-world data.

**Transition Type 1: Infection**

Infection in NetworkSIR occurs when an Infected node sends an agent to a Susceptible node. But, how does an infected node determine which and how many node(s) it should send an agent to if it has more than one edge? Furthermore, by what mechanism should agents be created?

To answer these questions, we make the following two assumptions:

1. The probability of an Infected individual ‘x’ infecting a Susceptible individual ‘y’ correlates linearly and positively with the duration of the CPI of x and y. Mathematically, this gives the following equation:

\[ P(\text{infection}) = k \cdot (\text{CPI}_{\text{duration}}) \quad k > 0 \]

This roughly equates to the assumption that if two individuals x and y are in close contact for an extended period, then it is more likely that infection will occur than if they were in contact for a brief period. This additionally implies that each second (or other time unit) of a CPI is as dangerous as any other.

2. Every Infected node uses the same mechanism that drives infection. This does not mean that each Infected node will always infect the same number of people, but rather, an algorithm which accurately describes the mechanism of infection for one Infected node (taking as input that node's CPIs/social relations) will be applicable for all Infected nodes. This has the important consequence that only one algorithm must be developed, rather than a new algorithm for every individual node.

From the above assumptions, we conclude that if we can determine a mechanism by which one can determine the duration of a CPI, we would be able to determine (relative to the individual) the probability that the individual will infect another. This would make it possible to determine the likelihood that an individual would infect a given person in their social circle as opposed to infecting another. However, first we must determine the mechanism which determines the length of a CPI, given that a CPI occurs. Mathematically, we must first find, for any individual i:

\[ P_i(\text{CPI with duration } t | \text{CPI}) \quad \forall i \]

We identified two cases\textsuperscript{4} by which we can calculate this probability: using a probability distribution given necessary parameters, or by approximating it using regression analysis. To generate these statistics, we used a contact network consisting of over 700 individuals and their CPIs, along with the duration of those CPIs, as collected from a U.S. high school\textsuperscript{10}.

We began by graphing the P(CPI with duration t | CPI) at both the individual and population level for the data set to determine by visual inspection which probability distributions might create an appropriate fit. We included the Binomial, Poisson, and Geometric distributions in our analysis. Of the continuous probability distributions, we included the Gaussian/Normal distribution, as certain formulations of the Central Limit Theorem applied to discrete probability distributions will yield a Gaussian/Normal distribution. Of the distributions, only the Geometric distribution appropriately fit the graphed data. We then applied Pearson’s Chi-Squared Test for Goodness of Fit\textsuperscript{18} with the following hypotheses:

1. H0: individual i’s CPI durations are Geometric Random distributed
2. H1: individual i’s CPI durations are not Geometric Random distributed

\textsuperscript{3} Due to page count limitations, we cannot elaborate in as much detail as we would like and still discuss all our results. For more details, refer to chapter 3.3 of “NetworkSIR and EnvironmentalSIR: Two Simple Distributed Mechanisms for Modeling Epidemics”.

\textsuperscript{4} We note that this analysis is mathematically insufficient to constitute a proof as there is a third possible case: the probability might be better approximated with some other method. However, it is infeasible to try the literally infinite number of alternatives (neural networks, etc.) and evaluate their performance with respect to each other.
All tested individuals rejected $H_0$ by a very large margin, with $P = 0.001$. Thus, we can say with high certainty that each individual's CPI durations are not Geometric Random distributed. Furthermore, since several probability distributions were tested for goodness of fit, and the Geometric Random distribution created the best fit yet was still insufficient, we conclude that it is unlikely that an individual's CPI durations can be expressed natively as a probability distribution.

Regression analysis provided a better fit. We performed linear, logarithmic, exponential, polynomial, and power series regression. Of these, the power series yielded the best fit with an $R^2$ value of 0.8636 (Figure 2). Figure 2 was generated by calculating the probability of each individual $i$ having a CPI of a given duration, given that $i$ would have a CPI. Each discrete CPI (and its probability of occurring) has its own point in Figure 2. We recorded 19,789 CPIs and their respective probabilities.

![Figure 2. The probability of a CPI of a given duration occurring](https://via.placeholder.com/150)

Since we have found an appropriate way of calculating $P_i$ (CPI with duration $x$ | CPI) $\forall i$, we can now fully propose a method of calculating $P(\text{infection})$. Knowing $P(\text{infection})$ can be used to assign relative probabilities that an Infected individual $i$ will infect a Susceptible individual $x$ as opposed to Susceptible individual $y$, given that $i$ has CPIs with both $x$ and $y$. These relative probabilities can then be scaled (normalized) such that their sum is equal to one for each individual. In the larger context of the NetworkSIR simulation, we can assign these relative probabilities to the edges of each node resulting in a probabilistic infection dynamics consistent with real-life behavior. We refer to this process as edge-weighting, and the steps are as follows:

1. Generate some probability $y$ using a Uniform Random Distribution
2. Use the equation given by regression analysis to calculate the associated CPI for that duration. Rounding to the nearest hundredths place and solving for $x$ (CPI duration) yields $x = \left(\frac{0.06}{y}\right)^{\frac{1}{17}}$
3. Weight the edge such that it is equal to the CPI duration calculated in step 2
4. Normalize all edge weights for a node such that they sum up to 1

When a node sends an agent, the edge weights above are used to determine who to send the agent to. To determine when to send an agent, we use the parameter $P(\text{Stay})$, equal to the probability that an infected node will not send an agent in a given round. It should be noted that non-pharmaceutical interventions such as handwashing and wearing of masks would alter $P(\text{Stay})$ by increasing it. Finally, to address the issue of when an Infected node should generate an agent, we require that an Infected node should always have the opportunity to infect other nodes. Therefore, only in the case that an Infected node begins a round without any agents will an agent be generated.

**Transition Type 2: Recovery/Removal**

We utilize a scalar value denoted as the Recovery Threshold to determine when a node should recover. Specifically, a node recovers if it has been infected for a number of rounds greater than the Recovery Threshold. This is motivated by the following:
Variations in recovery time are largely dependent on individual-specific factors (such as age, comorbidities, etc.), and we have previously assumed that these data are unknown. Further, individuals suffering death have a shorter period to “Recovery” than those who recuperate.

Randomness is already present in the model through our defined mechanism of infection.

Using a Recovery Threshold prevents nodes from remaining Infected for arbitrarily long amounts of time.

The parameter “Recovery Threshold” is naturally captured. It is a common practice to note how long individuals on average require to recover during symptom monitoring or how long it takes to die.

Readers interested in more implementation details and pseudocode should refer to chapter 3.4 of “NetworkSIR and EnvironmentalSIR: Two Simple Distributed Mechanisms for Modeling Epidemics”.

Evaluation

We evaluate our performance using two measures: the ability to 1. capture SIR dynamics, and 2. to predict the total number of individuals infected over the course of a pandemic with reasonable accuracy. To evaluate the former, we execute NetworkSIR with parameters designed to infect every individual: \( P(\text{stay}) = 0.8 \), Recovery Threshold = 30, and five initially Infected individuals. NetworkSIR is run over the same contact network used earlier which consisted of 789 individuals from a U.S. high school.  

!["Ideal" SIR Graph](image)

**Figure 3.** NetworkSIR graph showing asymptotic behavior

The results in Figure 3 highlights NetworkSIR’s major benefit over SIR. NetworkSIR produces a more symmetric infection curve, caused by eliminating SIR’s probabilistic mechanism of recovery/removal. A typical SIR graph’s infection curve tapers off more slowly, which equates to the assumption that as a pandemic continues individuals will take longer to recover or die. This assumption is not supported by facts. Thus, we feel comfortable claiming that NetworkSIR maintains a useful application of SIR dynamics in a network.

We now evaluate NetworkSIR’s prediction accuracy. We used statistics obtained from seasonal influenza to provide parameters for our simulation. The Recovery Threshold was determined by data from Harvard Health\(^1\) and the WHO\(^2\), which yielded an average recovery threshold of 5, and \( P(\text{stay}) \) was set to 0.78. NetworkSIR was then run for 500 simulations, and yielded an average of 16.7 percent of the population infected. Data from the NIH indicates that the worldwide average epidemic size is approximately 9% of a given population, while in the US it tends to be closer to 20%.

Data from the WHO\(^2\) similarly provide an estimate of 5-15%. Due to the fact that schools are notoriously densely populated (and thus should theoretically tend to the upper limit of percentage infected) we deem NetworkSIR’s prediction of 16.7% reasonably accurate.

Running NetworkSIR on a much smaller network consisting of 34 individuals and 78 edges\(^3\) for 500 times with the same \( P(\text{stay}) \) and Recovery Threshold parameters (but only one initially infectious individual) yielded an average epidemic size of 10.5%. We consider this a reasonable estimate, as it remained well within NIH and WHO predictions and is lower than the estimate obtained from the high school contact network—which represented a population with a higher population density. As NetworkSIR was able to provide reasonably accurate predictions
given two inherently different populations, we find these early results to be promising indicators of NetworkSIR’s potential for predicting a pandemic’s course.

EnvironmentalSIR

A key advantage of modeling disease spread explicitly through agents is that the effect of environment-to-human disease spread can be examined—particularly the effect of high-contact surfaces. To examine this, we modify NetworkSIR to have a new compartment, “Environment”. Environment nodes cannot transition to any other compartments, but can still infect nodes. To reflect the fact that most infectious material cannot survive outside of a living host, we define a “Sanitation Threshold” for Environment nodes. After ‘Sanitation Threshold’ number of rounds have passed, any agents residing within the Environment node are removed. By varying the length of the Sanitation Threshold, the effect of sanitation frequency (or, differences in the ability of infectious material to survive on surfaces) can be examined. A sample simulation of EnvironmentalSIR is shown in figure 4.

![Figure 4. A sample simulation of EnvironmentalSIR](image)

Although no further infection is possible, the simulation will continue for another round (until node B recovers). No further infection is possible as D is sanitized in Round 3, removing the agent that was passed to it in Round 2, and as before, only Infected nodes can generate agents.

To summarize, Environmental Nodes essentially perform two actions: infection and sanitization (i.e., the removal of agents after Sanitation Threshold number of rounds have passed). While Environmental Nodes cannot directly generate agents, we allow them to probabilistically pass any agents they might have to the nodes they have edges with using the same mechanism as Infected nodes. Environment Nodes in the context of the network represent high-contact surfaces that multiple individuals are regularly in contact with. As such, an edge between a node (representing an individual) and an Environment Node (perhaps representing a grocery store card-reader) represents the potential capacity of that individual to become infected through contact with the Environment Node.

Results

Due to the lack of prior work examining the effect of indirect disease spread in the context of the SIR model, we are unable to discuss the predictive ability of EnvironmentalSIR to capture “ideal” SIR dynamics. Instead, we provide comparisons between EnvironmentalSIR and NetworkSIR when parameterized with the same values. In particular, we evaluate the effect of varying the Sanitation Threshold and the number of Environment Nodes on the resulting epidemic dynamics.

As before, we run the simulation 500 times and report on the resulting statistics. We again use the same high school contact network, but this time randomly choose ten individuals (per run) to represent Environment Nodes. We set the Sanitation Threshold to be equal to two rounds. This resulted in a final average epidemic size of 14.2%, a reduction of 2.5%. Increasing the Sanitation Threshold to 5 rounds and holding all else constant yielded an average
epidemic size of 14.6%, a 2% increase. However, there were frequent cases in which small, secondary outbreaks occurred. Figure 5 depicts an execution of the simulation in which this occurred.

![EnvironmentalSIR Graph With Repeated Outbreaks](image)

**Figure 5.** Multiple outbreaks caused by the introduction of Environment Nodes with infrequent sanitation.

Further increasing the Sanitation Threshold to ten rounds resulted in a final average epidemic size of 15%. However, the number of secondary outbreaks increased as well. We therefore conclude that the primary effect of infrequent sanitation is an increase in an epidemic’s duration, and a secondary effect includes small increases in outbreak severity.

Increasing the number of Environment Nodes to 50 and setting the Sanitation Threshold equal to five rounds resulted in an average epidemic size of 12.1%. We therefore conclude that the primary effect of introducing environment nodes is a decrease on the final epidemic size. Introducing any environment nodes at all also seems to result in more asymmetric infection curves.

This reduction in overall epidemic size could be due to our method of representing them in the simulation: nodes representing individuals are changed to represent environmental fixtures. This could potentially lower the population density in key areas. Furthermore, Environment Nodes cannot produce agents—they can only remove them. This behavior may also contribute to smaller average epidemics.

**Limitations and Future Work**

The broadest class of limitations include those inherited from the SIR model, as it serves as the basis for NetworkSIR and consequently, EnvironmentalSIR. Additionally, both NetworkSIR and EnvironmentalSIR assume no knowledge about individuals within the population of interest. As such, features such as individual-level comorbidities, age, etc are ignored. Additionally ignored are any features other than CPI duration that could affect the likelihood of an individual being infected. Drawing on to NetworkSIR’s strength in its extensibility, future work can focus on incorporating more features into NetworkSIR’s predictions for increased accuracy. Example avenues of future work include replacing our proposed mechanism of edge-weighting with MLE (Maximum Likelihood Estimation) or MAP (Maximum a Posteriori) estimates for increased accuracy. However, in its current form, NetworkSIR is best used as a mechanism to study the flow of infectious material between individuals in a population, or as a framework for more complex models.

Additionally, both NetworkSIR and EnvironmentalSIR would benefit from further validation using real-world data. In particular, the collection of contact networks which include interactions between humans and commonly-visited environments is necessary for EnvironmentalSIR’s validation. A limitation which occurred from the lack of such networks is the re-use of the algorithm for infection (used by Environment Nodes) as used by Infected nodes in NetworkSIR. It is likely that humans do not interact with environmental fixtures in the same way that they do other humans, which would make this algorithm reuse inappropriate. However, due to lack of data, the development of a more appropriate mechanism is difficult.
Finally, the collection of large-scale datasets which track with certainty individual-level disease spread is necessary for the validation of both NetworkSIR and EnvironmentalSIR. Current datasets of this type are limited largely by size and method of collection (anecdotal accounts). However, it is likely that such data will be available in the near future due to the massively increased research interest in contact-tracing (automated or otherwise).

**Conclusion**

Accurate predictive modeling of pandemics allows for good public health decision making. To manage novel pandemics from a public health perspective, researchers and public health officials must be able to model disease dissemination to predict resource utilization such as bed capacity, staff requirements, and personal protective equipment. More importantly, as we learn how non-pharmaceutical interventions affect transmission risk, we must be able to quickly re-run the prediction models to adjust the forecast, which necessitates computationally efficient models. Furthermore, only accurate models that reflect the projected disease spread can help politicians and officials to determine large scale interventions such as mandatory mask wearing and shut downs of public spaces. 19

In this paper we offer two new models of disease dissemination: NetworkSIR and EnvironmentalSIR. Validation results and the higher verisimilitude of the models with actual disease spread suggest that they may offer better prediction. Using Environmental nodes to transmit agents allows for modeling of viral agents which are transmitted from surfaces and through ventilation. By building models that encompass environmental spread not linked to direct human-to-human contact, we can better explain pandemics. We encourage others to further test and validate our models.

**References**

18. Pearson, Karl. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Philosophical Magazine 1900; Series 5. 50 (302): 157–175.
36. Davis., K. Table of Chi-square Statistics [Internet], University of Texas; 2020. Available from: https://web.ma.utexas.edu/users/davis/375/popecol/tables/chisq.html
Epilepsy-Connect: An Integrated Knowledgebase for Characterizing Alterations in Consciousness State of Pharmacoresistant Epilepsy Patients

Katrina Prantzalos, MS¹, Jianzhe Zhang, MS¹, Nassim Shafiabadi, MD¹,², Guadalupe Fernandez- BacaVaca, MD², Satya S. Sahoo, PhD¹,²

¹ Department of Population and Quantitative Health Sciences, Case Western Reserve University School of Medicine, Cleveland, OH, USA
² Department of Neurology, University Hospitals Cleveland Medical Center, Cleveland, OH, USA

Abstract

Alterations in consciousness state are a defining characteristic of focal epileptic seizures. Consequently, understanding the complex changes in neurocognitive networks which underpin seizure-induced alterations in consciousness state is important for advancement in seizure classification. Comprehension of these changes are complicated by a lack of data standardization; however, the use of a common terminological system or ontology in a patient registry minimizes this issue. In this paper, we introduce an integrated knowledgebase called Epilepsy-Connect to improve the understanding of changes in consciousness states during focal seizures of pharmacoresistant epilepsy patients. This registry catalogues over 809 seizures from 70 patients at University Hospital’s Epilepsy Center who were undergoing stereotactic electroencephalography (SEEG) monitoring as part of an evaluation for surgical intervention. Although Epilepsy-Connect focuses on consciousness states, it aims to enable users to leverage data from an informatics platform to analyze epilepsy data in a streamlined manner.

Epilepsy-Connect is available at https://bmhinformatics.case.edu/Epilepsyconnect/login/.

1. Introduction

Epilepsy is a serious neurological disease that is defined as having at least two unprovoked seizures in a 24-hour time period and is the fourth leading non-communicable disease impacting global years lived with disability (YLD)[1], [2]. More than 52 million people worldwide are affected by epilepsy, resulting in 18.3 million global YLD[2]. Approximately one third of these people suffer from refractory or drug-resistant epilepsy, leaving them with uncontrolled seizures[3].

Epilepsy is a spectrum condition that broadly encompasses all seizure disorders, with alterations of consciousness during seizure events being a core component of epileptic seizures[4]. While successful treatment of epilepsy requires refined classification of the disorder, informative classification guidelines that focus on changes in consciousness state have been difficult to achieve. In 2017, the International League Against Epilepsy (ILAE) updated the definition of epilepsy to include several new focal and generalized seizure types[5]. Despite offering more granular classifiers such as awareness, these ILAE seizure classifications have been criticized for a lack of clarity and brevity[6]–[8]. The “Four-dimensional epilepsy classification” is an approach that aims to classify seizures based exclusively on ictal symptomatology, and it includes precise and detailed descriptions of the nature of changes in consciousness state[7], [9]. This proposed approach to use clinical features to better delineate focal seizures reduces the dependency on an abnormal, correctly interpreted, available electroencephalograph (EEG)[8].

A better understanding of the changes in consciousness state during seizures is essential for both improvement in treatment of epilepsy as well as potentially enhancing the safety of epilepsy patients during seizure events [4], [8], [10], [11]. In addition to limitations of existing classification systems with respect to consciousness state during epileptic seizures, the lack of standardization in case ascertainment, data collection methods, and data reporting methods also impedes advancement in epilepsy research related to consciousness. In this paper, we propose that the design and implementation of a patient registry focused on alterations of consciousness state in epilepsy using a common terminological system or ontology can address many of the data management challenges faced by clinical researchers [12]–[16]. Ontologies are knowledge reference models that often use description logic to represent the relationships between concepts or terms in a way that allows computational logical reasoning, which allows for mass integration and analysis of large amounts of heterogeneous data; this logic is widely used to enhance patient classification and other data analysis tasks [14], [16]. Therefore, ontologies can play a key role in managing multidimensional data from different sources in a consciousness focused patient registry.
Patient registries have become a common way to facilitate sharing comprehensive, diagnostic, condition-specific data in order to enhance clinical and scientific research [12]. There are several existing registries that cover various aspects of epilepsy, each providing unique insights into different aspects of therapies, rare variants, and general health related to epilepsy (Table 1). To the best of our knowledge, there is no existing registry that focuses on changes in consciousness states in epilepsy patients and the use of the registry data for computational analysis of the patient data to study changes in brain network topology during seizures. Furthermore, most existing epilepsy registries do not include publicly accessible data, making it difficult for researchers to perform exploratory data analysis or relate these highly specific findings to more general topics. To address this significant gap in the study of consciousness states in epilepsy patients, we introduce an integrated knowledgebase called Epilepsy-Connect that uses ontology for standardization of data in a new patient registry to study changes in consciousness states during epileptic seizures (https://bmhinformatics.case.edu/Epilepsyconnect). The broader objective of the Epilepsy-Connect informatics platform is to enable a greater understanding of the consciousness state system, ultimately improving diagnostic accuracy, patient care, and the prediction of surgical necessity and/or outcomes.

Table 1. An example of epilepsy registries already in existence and their respective purposes.

<table>
<thead>
<tr>
<th>Registry Name</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rare Epilepsy Network Registry (REN) [17]</td>
<td>To catalogue and understand 32 different rare epilepsies with the goal improving treatments and quality of life [17]</td>
</tr>
<tr>
<td>North American SUDEP Registry (NASR) [18]</td>
<td>To obtain medical records, family interviews, and other reports of cases of sudden unexpected death in epilepsy [18]</td>
</tr>
<tr>
<td>Stockholm Incidence Registry of Epilepsy (SIRE) [19]</td>
<td>To identify prospective epilepsy cases in patients with newly diagnosed single unprovoked seizures [19]</td>
</tr>
<tr>
<td>International Ion Channel Epilepsy Patient Registry (IICEPR) [20]</td>
<td>To collect data on patients with epilepsy caused by an ion channel mutation [20]</td>
</tr>
<tr>
<td>Medtronic Registry for Epilepsy (MORE) [21]</td>
<td>To evaluate the long-term effects of deep brain stimulation for the treatment of refractory epilepsy [21]</td>
</tr>
<tr>
<td>International Registry of Antiepileptic Drugs and Pregnancy (EURAP) [22]</td>
<td>To determine risks associated with antiepileptic drug use during pregnancy [22]</td>
</tr>
<tr>
<td>Epilepsy Birth Control Registry (EBCR) [23]</td>
<td>To evaluate the contraceptive practices of women with epilepsy [23]</td>
</tr>
</tbody>
</table>

1.1 Background

**Significance of Consciousness State in Epilepsy Neurological Disorder.** Up to 70% of focal epilepsy patients do not have freedom from seizures despite optimal drug therapy [3]. However, surgery to resect the epileptogenic zone often does not result in seizure-free patients, and they continue to experience significant deterioration in their quality of life due to repeated seizures, including loss or change in consciousness state that endangers their safety [11], [24]. The analysis of changes in consciousness states during recurrent seizures has been suggested as a potential method to improve evaluation of patients for surgery [3], [24]. Consciousness is a comprehensive yet complicated term lacking one concrete definition [4], [25]. Consciousness, while not concretely defined, is believed to depend upon continuous integration and processing of information [4]. Interruptions in this continuous integration and processing of information—often presenting as changes in awareness, attention, arousal, responsiveness, ability to interact, and/or memory—are noted as alterations in consciousness state [4], [11]. Focal seizures span various degrees of impairments or alterations of consciousness, including: (1) no impairment, (2) observable motor or autonomic components, (3) sensory or psychic auras, (4) dyscognition, (5) dialepsis, (6) ictal delirium, (8) evolution to a bilateral, convulsive seizure, and (9) epileptic coma [4], [26].

Losses of, or alterations in, consciousness state in patients with epilepsy are complex phenomena that involve changes...
in neurocognitive networks such as the default mode network (DMN), salience network (SN), and central executive network (CEN)[27]–[30]. These impairments may be identified by various neurological characteristics and can be measured objectively by assessing a patient’s amnesia of events occurring during epileptic seizures and their responsiveness to external stimuli during epileptic seizure [4], [11]. Several studies have been conducted to link changes in consciousness to specific brain regions[11], [31], [32] and to changes in brain activity in several cortical regions and subcortical arousal systems[11], [27]–[30], [33], [34]. These studies, however, fail to distinguish between the various states of consciousness. Each state of consciousness has its own unique clinical presentation, and it has been noted that each state of consciousness has unique characteristics during a seizure. We propose to analyze signal recordings from refractory patients to characterize changes in brain network topology and correlate the changes to alterations in consciousness during seizure events using de-identified patient data from the Epilepsy Center at the University Hospitals Cleveland Medical Center (UH-CMC).

Epilepsy-Connect Patient Registry Data. The UH-CMC is a Tier 4 epilepsy center that conducts research on the causes and therapies of epilepsy with refractory epilepsy patients receiving care in the epilepsy monitoring unit (EMU). In the EMU, patients undergo continuous monitoring of brain activity for five days, including continuous video monitoring. During this period, anti-seizure medications are temporarily reduced or stopped in order to safely induce medically supervised seizures so that epileptologists may determine the epileptogenic zone and, by extension, determine the best course of therapy for the patient. Patients are monitored using intracranial electrodes that record stereotactic electroencephalography (SEEG) to localize seizure foci. We use de-identified patient data from the UH-CMC with Institutional Review Board (IRB) approval to populate the Epilepsy-Connect patient registry.

Network Analysis of SEEG Data for Characterizing Alterations of Consciousness State in Epilepsy Patients. It is well understood that focal seizures originate within networks where initial activation is limited to one cerebral hemisphere[26], [35]. Current focal epilepsy models acknowledge that the epileptogenic zone is a distributed network involving inter-regional anatomo-functional relationships among brain areas[24], [36]. These networks have not been fully characterized yet, and research for more accurate prediction of surgical intervention outcomes using these networks is ongoing[24], [36].

One increasingly popular body of research has emphasized the utility of network analysis methods that model epileptic seizure networks as graph models consisting of brain locations as nodes and the interaction between them as edges in a graph to understand each of these potential network mechanisms[37]. Network topology has been used to study general changes in ictal network properties spanning the duration of the seizure[38], [39] or at seizure onset[40]. Research has even considered changes in interictal network topology[37], [38], [41], [42]. While there is an abundance of literature focused on developing general understandings of epilepsy networks, little is known about the relationship between these networks and the differences in the various impaired states of consciousness.

The Epilepsy-Connect knowledgebase aims to serve as an integrated neuroinformatics resource with a patient registry and network analysis tools, which can be directly applied to the patient data. The Epilepsy-Connect knowledgebase features a query interface that allows users to create patient cohorts based on specific inclusion/exclusion criteria. In the next section, we describe the implementation details of the Epilepsy-Connect knowledgebase, including the workflow used to populate the registry, the development of the informatics component, and the functionalities currently supported by this integrated neuroinformatics tool.

2. Methods

The Epilepsy-Connect knowledgebase was developed in close collaboration with clinicians at the UH-CMC to support clinical research together with patient care. Figure 1 shows the workflow used in the development of the Epilepsy-Connect knowledgebase. The three primary objectives of the Epilepsy-Connect knowledgebase are:
1. Support the creation of a patient registry using standardized terminology modeled in a widely used epilepsy-focused ontology [16];

2. Enable researchers to perform patient-cohort queries over the patient registry data for hypothesis development and validation related to alteration of consciousness state in pharmacoresistant epilepsy patients who undergo evaluation for surgical evaluation; and

3. Support users to manage their cohort queries and query results, which can be used for data sharing as well as dissemination of their study results.

2.1 The architecture and development of the Epilepsy-Connect platform. The Epilepsy-Connect platform consists of a set of modules that support: (1) creation of a cohort query using a visual query interface, and (2) management of user information, including query history. The current version of Epilepsy-Connect has been developed using the Django web application framework, which uses the Python programming language and features a large number of libraries and modules that support variety of data processing and analysis tasks including access to ontology files. The architecture of Epilepsy-Connect allows it to integrate and manage data from the UH-CMC using software modules for data pre-processing and conversion to a common data model following the well-known Extract Transform Load (ETL) approach. We note that as part of the data processing step, multiple terms in the study data conform to the ontology classes in the Epilepsy Ontology, which facilitates the standardization of the data terms in the patient registry module of the platform.

The Epilepsy-Connect platform uses the Model View Template (MVT) approach with data in the registry component managed using an object relational data model, the user interface is managed by the View component, and the user interaction with various features of the software is mediated by the Template, which conforms to the Django framework. The Epilepsy-Connect platform is accessed via a web browser with role-based access control (RBAC) with users assigned to a user group. The access to data in the registry component is governed by the UH-CMC IRB with a data user agreement (DUA) required for download of the data. Users follow an intuitive query process to create a study cohort using a visual query composition process (results are discussed in Section 3).

2.2 Study Data. At present, the Epilepsy-Connect registry hosts de-identified retrospective patient data from UH-CMC who were evaluated for surgical intervention. The preoperative evaluation data were gathered from each patient’s discharge summary following invasive EEG/video evaluation. This data was split into patient demographics, medical history, seizure etiology, medical imaging records and findings, neuropsychological testing information, and clinically observed representations of patient alterations of consciousness. Further, de-identified SEEG recordings for each patient are processed using an in-house signal processing workflow called the NeuroIntegrative Connectivity (NIC) tool to segment and transform signal data stored in the European Data Format (EDF) to a Javascript Object Notation (JSON) format called Cloudwave Signal Format (CSF) [43], [44]. The SEEG recordings are analyzed by the UH-CMC clinicians during EEG reading sessions to determine seizure details and define ictal and interictal periods. Further, clinicians also note the specific electrode contact involved in seizure events, which are used as input parameters for the computational neuroscience workflow used in this project to characterize topological changes associated with change in consciousness state of these patients. The seizure details, ictal time periods, and the list of electrode contacts together with other relevant patient details are stored in the
Epilepsy-Connect knowledge as part of the ETL process.

Summary statistics of the data were calculated in R (version 4.0.4). A total of 70 patients, ranging from the age of 15 to 69 (median age of 35), were included in the registry. Table 2 shows a summary of the patient demographics. The number of males and females included was approximately equal, and age did not differ significantly between the sexes ($p = 0.97$, $\chi^2 = 0.001$). A total of 809 seizures were analyzed for the registry. The seizures for one patient were excluded from this calculation since the patient was reported to have had “one seizure every two minutes” over multiple days within the study. Of these seventy participants, sixty-one reported having at least one seizure with an alteration of consciousness. Forty-two patients were reported to have had an aura during at least one seizure; twenty-seven patients experienced dialepsis during at least one seizure; six patients presented with dyscognition in at least one seizure; and thirty-one patients had at least one seizure without any alterations of consciousness. Access to the patient registry data through the Epilepsy-Connect query interface is available as a “guest user” and a “registered user”. A registered user account will be created after the user has completed the data user agreement with IRB approval. A registered user has full access to the Epilepsy-Connect knowledgebase; however, a “guest user” can explore a restricted set of data elements through the user interface.

2.3. Epilepsy Ontology and Standardization of Terms. The Epilepsy-Connect uses the Epilepsy Ontology (listed in the National Center for Biomedical Ontologies as the Epilepsy and Seizure Ontology, EpSO) as a reference ontology to standardize the storage and querying of data [16]. The Epilepsy Ontology has been iteratively developed using input from multiple epilepsy domain experts as part of the ILAE Big Data task force to comprehensively model various aspects of epilepsy and seizures, including genetics, neuropathology, medication, and brain anatomy. Epilepsy Ontology uses the description-logic-based Web Ontology Language (OWL) to construct a class hierarchy, link together the ontology classes using properties, and apply sophisticated class-level restrictions to model epilepsy information at a fine-level of granularity. The Epilepsy Ontology uses ontology-engineering best practices to re-use existing ontology terms from the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT), the Gene Ontology, RxNorm for drug information, and the Foundational Model of Anatomy (FMA) for modeling brain regions. A phased approach was used to extend the Epilepsy Ontology for use in the Epilepsy-Connect project.

In the first phase, we enumerated all the data elements in the Epilepsy-Connect registry, which can be broadly classified into six categories of patient evaluation results (e.g., etiology, neurological review, and family history); results of neurological exam (e.g., memory, attention, motor, and reflexes); details of imaging procedures (e.g., imaging modality and status); seizure events (e.g., start and end of seizure events, onset and end frequency); alterations of consciousness state (e.g., aura, dialeptic, or apraxia); and details of alterations of consciousness events (e.g., post-seizure memory retention, fine distal movement). In the next phase, we modeled these data elements in the Epilepsy Ontology, for example Figure 2 shows a part of the class hierarchy describing EEG patterns. These ontology terms are being used in the Epilepsy-Connect project for both annotation of database as well as to support the functionalities of the user interface. In particular, we are integrating the Epilepsy Ontology in the user interface to support patient cohort queries using the ontology-based database access (OBDA) framework. The use of OBDA will allow users to leverage semantic matching of query terms to information in the patient registry without being constrained by string matching only. At present, the Epilepsy-Connect query interface uses the ontology terms to create a customized database for supporting the query composition task using drop-down menu values. The use of the Epilepsy Ontology in the patient registry will also enable easier sharing of data using common terminology for data annotation.

<table>
<thead>
<tr>
<th>Table 2. Summary statistics for participants</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Percent of People (n = 70)</strong></td>
</tr>
<tr>
<td><strong>Mean Age (Standard Deviation)</strong></td>
</tr>
<tr>
<td>Mean Age</td>
</tr>
<tr>
<td><strong>% Male</strong></td>
</tr>
<tr>
<td><strong>% Right-Handed</strong></td>
</tr>
<tr>
<td><strong>Epileptogenic Zone (%)</strong></td>
</tr>
<tr>
<td>Bitemporal</td>
</tr>
<tr>
<td>Left Hemisphere</td>
</tr>
<tr>
<td>Right Hemisphere</td>
</tr>
<tr>
<td><strong>Alteration of Consciousness (%)</strong></td>
</tr>
<tr>
<td>Aura</td>
</tr>
<tr>
<td>Dialepsis</td>
</tr>
<tr>
<td>Dyscognition</td>
</tr>
<tr>
<td>None</td>
</tr>
</tbody>
</table>
Network Analysis of SEEG Data using Patient Cohort Query. In addition to supporting patient cohort queries, the Epilepsy-Connect knowledgebase is being integrated with a network analysis tool called the Neuro-Integrative Connectivity (NIC) platform, which was developed in our prior work[43]. The NIC platform converts SEEG European Data Format (EDF) files into Cloudwave Signal Format (CSF) files, a JavaScript Object Notation (JSON) based human-readable format previously developed by our lab[44]. Once files have been converted, the NIC tool may be used to investigate statistical correlations of brain functional connectivity. Users are able to choose between Pearson Correlation[45], Mean Phase Coherence[46], and Non-linear correlation[47] for their measurement(s) of correlation. Furthermore, the NIC workflow allows users to analyze the change in network graph topology throughout seizures via methods in algebraic topology. Outputs include the clustering coefficient, Betti numbers, characteristic path length, and global efficiency. The NIC tool is available for download at https://bmhinformatics.case.edu/nicworkflow/accounts/login/.

After the user creates a study cohort using the Epilepsy-Connect user interface, the NIC tool will be able to be invoked to process and compute network measures over the SEEG data. At present, we created a cohort with patients with two seizures for further analysis. The corresponding SEEG files were processed using the NIC platform to characterize the network characteristics of the epileptic seizures, including global and local connectivity with corresponding disruption of the small-world network topology, which is a characteristic of brain network in healthy subjects. In the next section, we describe the results generated by the Epilepsy-Connect knowledgebase via the user interface.

3. Results

Epilepsy-Connect User Interface. The Epilepsy-Connect user interface consists of an intuitive query composition module (Figure 3), which supports both drop-down menu values as well as a “Drag-and-drop” feature. Using a stepwise process, users can select at least one variable from a list of curated patient cohort variables. Variables provide a drop-down list of factors which can be used to impose search criteria limits; numeric variables will provide sliders whereas qualitative responses will allow for radio button selection. At this time, queries are limited to “OR” logic. Once a user has completed a query, they will be directed to the results page to view their results. Restricted access users will have the option at this point to download the resulting dataset in a comma separated values (CSV) format. For an example of functionality, we have chosen to illustrate the query criteria that a patient had at least one dialeptic seizure (Figure 4).

Additional features have been implemented for restricted access users. Users with a registered account will be able to review the data in the registry as a whole by visiting the “data” tab in the upper right-hand corner of the page. Furthermore, they will be able to revisit their prior queries, by utilizing the “history” tab in the upper right-hand corner of the page.
4. Discussion & Limitations

We created an integrated knowledgebase called Epilepsy-Connect to improve the understanding of changes in consciousness states during focal seizures of pharmacoresistant epilepsy patients. This registry catalogues over 809 seizures from 70 patients at the Epilepsy Monitoring Unit within University Hospital’s Epilepsy Center who were undergoing SEEG monitoring as part of an evaluation for surgical intervention. Public use data for the registry includes information on general patient demographics, clinical symptoms that patients experienced during at least one of their seizures, and the alterations of consciousness that patients experienced during at least one of their seizures. We propose to deploy Epilepsy-Connect in the EMU of UH-CMC to support clinical research and improve patient care.

The broader objective of the Epilepsy-Connect informatics platform is to enable greater understanding of the consciousness state system that could be useful in improving diagnostic accuracy and patient care, as well as predicting surgical necessity and/or outcomes. To demonstrate and validate the use of Epilepsy-Connect for this purpose, we composed a simple query of the public access data to examine patients who had experienced at least one diaphaptic seizure during the recorded evaluation. The resulting cohort included six patients. While the results of this case study are enlightening, further analysis is needed to fully comprehend the complexities of the various networks associated with alterations of consciousness states. Comparisons across the selected cohort could reveal network changes that occur during seizures in which a patient presents with diaphapsis. Moreover, comparisons across multiple cohorts could foster understanding of network changes that occur during any seizure-induced alteration of consciousness state.

Although Epilepsy-Connect was designed with a focus on consciousness states, it includes a large range of data that can be utilized to explore epilepsy from several perspectives. Epilepsy-Connect’s standardized and comparable data collection makes it ideal for sharing comprehensive diagnostic data in order to improve research into epilepsy treatment methods and new intervention techniques, while still providing unique insights into seizure-induced alterations of consciousness states[12]-[16]. In an effort to make Epilepsy-Connect more universal, future releases are expected to add information such as weighted diffusion tensor imaging results for further connectome analysis, patient surgery decisions and outcomes, and additional patient neuropsychological testing. Furthermore, the modular architecture of the registry allows for queries to be easily modified, thus streamlining data collection procedures for any interested investigators.
5. Conclusions

Alterations of consciousness state are important in defining epileptic seizures[6]–[8], [10], and developing comprehension of seizure-induced changes to consciousness state is essential for improving seizure classification[8], [10]. One aspect of this is to understand the complex changes in neurocognitive networks which underpin alterations of consciousness states during epileptic seizures[27]–[30]. However, this is not easy, as the lack of standardization in case ascertainment, data collection methods, and data reporting methods tend to make data difficult to compare or validate across studies[12]–[15]. Patient registries work to minimize these issues[12]. In this paper, we introduced an integrated knowledgebase called Epilepsy-Connect that supports standardized and comparable data collection of patients with refractory focal epilepsy, and showed one potential use case for improving the understanding of changes in consciousness states during epileptic seizures. Although Epilepsy-Connect was designed with a focus on consciousness states, it aims to enable users to leverage data from an informatics platform to analyze epilepsy data in a streamlined manner in order to bolster scientific advancement.

Acknowledgements

This project was supported in part by the Clinical and Translational Science Collaborative (CTSC) of Cleveland which is funded by the National Institutes of Health (NIH), National Center for Advancing Translational Science (NCATS), Clinical and Translational Science Award (CTSA) grant, UL1TR002548. The content is solely the responsibility of the authors and do not necessarily represent the official views of the NIH. This project was also supported in part by the NSF grant #1636850.
References


[19] University of Michigan Hospital and Dravet Syndrome Foundation & Ion Channel Epilepsy Alliance, “International Ion Channel Epilepsy Patient Registry.” https://sites.google.com/umich.edu/iicepr


Investigating Health Information Technology Usage by Sociodemographic Subpopulations to Increase Community Engagement in Healthcare: An Analysis of the Health Information National Trends Survey

Geetanjali Rajamani, BS1, Lianne Kurina, PhD1,2, Lisa Goldman Rosas, PhD, MPH2
1Department of Human Biology and 2School of Medicine, Stanford University, CA

Abstract

It is well known that the US is plagued by health inequities: unjust differences in morbidity and mortality rates by sociodemographic factors. A potential method to address such inequities lies in utilizing health information technologies (HIT) to reach under-resourced populations and increase their involvement in healthcare. Previous researchers have done just this, using HIT tools to engage under-resourced communities and improve outcomes. However, it is unclear how HIT usage varies by sociodemographic characteristics. This study investigated this question through analysis of the Health Information National Trends Survey (HINTS) and proposed tailored HIT interventions for specific subpopulations. Internet, smartphone, and wearable device usage were analyzed by age, race/ethnicity, educational attainment, and income; purposes of HIT usage were assessed; and logistic regression models were conducted to determine associations between purposes of HIT usage and sociodemographic predictors. Results showed that Black/African American, Latinx, and Asian populations all had significantly increased use of health videos, while participants with lower educational attainment had significantly decreased use of many HIT tools. Thus, this study highlights effective interventions for specific racial/ethnic populations and showcases a need for HIT tools inclusive towards low education populations to increase their engagement in healthcare and reduce inequities.

Introduction

There exist many health inequities in the United States – significant differences in morbidity and mortality rates by social factors such as race, socioeconomic status (SES), and educational attainment. For example, African Americans make up 13% of the US population but account for 42% of new HIV diagnoses, and Latinx make up 18% of the population and account for 27% of new diagnoses. American Indian/Alaska Natives (AIAN) are twice as likely to have diabetes as whites. In the COVID-19 pandemic, minorities disproportionately bear the burden of COVID-19 cases, hospitalizations, and deaths. Such inequities arise from a complex combination of factors, including lack of access to adequate/affordable care, low health literacy rates, racism, and various systemic injustices. Solutions to address these inequities are of the utmost importance, not only to improve the health of the nation, but also to serve social justice in healthcare.

In today’s digital world, technology is becoming rapidly integrated with many aspects of life, including healthcare. Health information technology (HIT) is a prominent field; a variety of apps exist to help with disease management; the electronic health record (EHR) collects health data from thousands of patients electronically; and even virtual reality is being used to treat mental health disorders. Thus, a potential solution to addressing health inequities may also be found in the realm of HIT. Specifically, HIT can be utilized to reach under-resourced populations, provide them with access to health information and health care, and reduce suboptimal outcomes for these populations.

Review of Relevant Research

Previous research has shown that HIT can be used as a form of community engagement to involve sets of under-resourced communities, rather than just individuals, to be invested in their own healthcare and is a powerful strategy to reduce health inequities. For example, a review study by Kranz et al. (2018) found that when safety net clinics – serving under-resourced populations – utilized HIT for coordinating appointments/transportation with patients, facilitating communication with providers, and a host of other purposes, patient outcomes such as hemoglobin A1c, maternal and child health, and adult screening practices improved.

However, there is much room for expansion in the literature regarding how HIT usage varies by demographic characteristics. A study by Mackert et al. (2016) examined how HIT usage varies by health literacy of participants. They found that those with an “adequate” health literacy level or higher were more likely to use fitness apps, nutrition apps, activity trackers, and patient portals than those with a “less than adequate” health literacy level, suggesting that current HIT tools must be modified in order to cater to low health literacy communities. A study by Wang et al. (2019) analyzed HIT usage patterns of older adults (average age 70 years), finding that although few are currently using...
mobile technologies, many are open to the idea. Overall, these studies show that although researchers have examined how certain subpopulations utilize HIT – such as health literate adults and older individuals – few studies exist regarding how HIT usage generally varies across a variety of subgroups.

**Project Aims**

It is unknown exactly what sociodemographic factors are correlated with HIT usage – i.e. do some groups use smartphone-based technology more and do others use web-based technology more? Having a better understanding of what sociodemographic factors are correlated with HIT usage will help better tailor studies regarding which modality of technology should be used to engage a specific group of people. This increased engagement can potentially lead to increased prevention of chronic diseases, increased screening rates in under-resourced communities, and overall reduced inequities and improved health.

This study aims to address this gap by analyzing whether and how sociodemographic factors such as race/ethnicity, education, and total income are associated with HIT-related usage factors, including access to internet, smartphone usage, use of an electronic wearable device, and more. The specific questions investigated were:

- How do rates of smartphone, internet, and wearable device usage vary across sociodemographic groups (race/ethnicity, income, education)?
- What are the methods in which HIT is used in healthcare to help with decisions and communications (e.g. texts with providers, health apps, health videos, etc.)?
- Are there differences in use of health and wellness apps by sociodemographic status (race/ethnicity, income, education)?

This study will thus help understand what modalities of HIT are best for which sociodemographic groups (and thus guide future interventions) and provide a better picture of the role of technology as a potential strategy to address health inequities.

**Methods**

To determine how HIT usage varies by sociodemographic factors, the Health Information National Trends Survey (HINTS), sponsored by the National Cancer Institute (NCI), was utilized.

**Sample**

The HINTS survey was developed by the NCI to better understand the rapidly evolving field of health communication, which includes HIT. HINTS is a cross-sectional survey of a nationally representative sample of American adults, administered regularly since 2003, and HINTS data are publicly available. This survey instrument consists of a total of 129 close-ended questions organized into 14 sections, and all data are quantitative. The data are de-identified and made available for public use by adhering to standards for analysis and reporting. The target population is adults age 18 years and above. For this study, HINTS 5 (Cycle 3) was utilized, collected from January-May 2019 (N=5438). Detailed sampling strategies and survey design methodologies are available in the HINTS 5, Cycle 3 methodology report.

**Measures: Sociodemographic Factors**

The following sociodemographic variables, taken from Section O of the HINTS questionnaire, were included in analyses: age; highest grade or level of schooling; Hispanic/Latino/Spanish origin; race; annual income. For the purposes of this study, the race and ethnicity variables were combined and are presented as Non-Hispanic White, Non-Hispanic Black, Hispanic/Latinx, Asian and Other (including AIAN, Pacific Islander, and multiple races). Educational attainment categories were also combined and are presented as: high school degree or less, some post-high school education, college degree and postgraduate studies.

**Measures: HIT Usage**

Table 1 lists the HIT usage variables (section B) of the HINTS questionnaire that were included in analyses, along with their response options. A total of 11 questions were included, of which 3 questions pertain to internet use, 5 inquire about smartphone ownership and usage, and 3 relate to electronic wearable device use.
Table 1: Selected HIT Usage Variables, HINTS 5 (Cycle 3)

<table>
<thead>
<tr>
<th>HIT Usage Variables</th>
<th>Response Options</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Internet Use</strong></td>
<td></td>
</tr>
<tr>
<td>B1: Do you ever go on-line to access the Internet or World Wide Web, or to send and receive e-mail?</td>
<td>Yes; No</td>
</tr>
<tr>
<td>B3: In the past 12 months, have you used the Internet to look for information about cancer for yourself?</td>
<td>Yes; No</td>
</tr>
<tr>
<td>B14: Sometimes people use the Internet to connect with other people online through social networks like Facebook or Twitter. This is often called “social media.” In the past 12 months, have you used the Internet to: share health information on social networking sites, such as Facebook or Twitter?; participate in an online forum or support group for people with a similar health or medical issues?; watch health-related videos on YouTube?</td>
<td>Yes; No</td>
</tr>
<tr>
<td><strong>Smartphone Use</strong></td>
<td></td>
</tr>
<tr>
<td>B5: In the past 12 months, have you used a computer, smartphone, or other electronic means to: look for health or medical information for yourself; buy medicine or vitamins online; use e-mail or the Internet to communicate with a doctor or a doctor’s office; track health care charges and costs; look up medical test results; make appointments with a health care provider; look for information about the harms of electronic or e-cigarettes?</td>
<td>Yes; No</td>
</tr>
<tr>
<td>B6: Please indicate if you have a smartphone.</td>
<td>Check box if yes</td>
</tr>
<tr>
<td>B7: On your smartphone, do you have any “apps” related to health and wellness?</td>
<td>Yes; No; Don’t know</td>
</tr>
<tr>
<td>B8: Has your smartphone helped you: track progress on a health-related goal such as quitting smoking, losing weight, or increasing physical activity?: make a decision about how to treat an illness or condition?: facilitate discussions with your health care provider?</td>
<td>Yes; No</td>
</tr>
<tr>
<td>B15: have you sent a text message to or received a text message from a doctor or other health care professional within the last 12 months?</td>
<td>Yes; No; Don’t know</td>
</tr>
<tr>
<td><strong>Electronic Wearable Device Usage</strong></td>
<td></td>
</tr>
<tr>
<td>B9: In the past 12 months, have you used an electronic wearable device to monitor or track your health or activity? For example, a Fitbit, Apple Watch, or Garmin Vivofit.</td>
<td>Yes; No</td>
</tr>
<tr>
<td>B10: In the past month, how often did you use a wearable device to track your health?</td>
<td>Every day; Almost every day; 1-2 times per week; Less than once per week; I did not use a wearable device in the past month</td>
</tr>
<tr>
<td>B11: Would you be willing to share health data from your wearable device with your health care provider? Your family or friends?</td>
<td>Yes; No</td>
</tr>
</tbody>
</table>

Data Analysis

Data analysis was conducted using the software Stata/IC 16.1 (StataCorp LLC, TX, USA). First, univariate analyses were conducted to understand the purposes of HIT usage. Among those who indicated they were internet users, analyses were performed to determine how many used internet for: looking up cancer information; sharing health information on social media; participating in online health forums; and watching YouTube health videos (Figure 1). Among those who indicated they were smartphone users, analyses were performed to determine how many used smartphones for: looking up health information; buying medicines online; communicating with health care providers; tracking health care costs; looking up medical test results; making health care appointments; looking up information regarding harms of e-cigarettes; using health/wellness apps; tracking progress on health-related goals; making decisions on treatment plans; and sending/receiving texts from providers (Figure 2). Finally, among those who indicated they were electronic wearable device users, analyses were performed to determine how many use these devices: every day; almost every day; 1-2 times per week; less than once per week; not in the past month; and how many would be willing to share data with providers or family/friends (Figure 3).
Next, bivariate analyses were conducted to examine the relationship between HIT usage and sociodemographic variables. Whether a participant uses internet or not was analyzed by age, a combined ethnicity/race variable, education, and income categories (Table 2). Similarly, smartphone usage was analyzed by the same demographic variables (Table 3), as was electronic wearable device usage (Table 4). Chi-squared tests and two-tailed t-tests were conducted to determine the significance of these various relationships.

Finally, four separate multivariate logistic regression models were conducted to understand how some purposes of HIT usage (mentioned in Figures 1-3) vary by sociodemographic predictors. Specifically, smartphone use for engaging with health and wellness apps, internet use for watching YouTube health videos, smartphones to look up health information, and smartphones to communicate with providers were each examined separately, with age, ethnicity/race, education, and income as predictors (Table 5). These four variables were chosen for the regression models because they were the most commonly reported forms of HIT usage.

Results

Sociodemographic Analysis of Internet Users & Purposes of Usage

Table 2 summarizes the sociodemographic factors of participants who reported using internet (N=4322) versus those who reported not using internet (N=1074). The average age of internet users was 54.1 years (SD 16.4), while the average age of non-internet users was 68.6 years (SD 13.7) – a significant difference (p<0.0001). More non-Hispanic whites and Asians report using internet compared to non-Hispanic Blacks and Hispanic/Latinx populations; more educated people report using internet compared to less educated populations; and more people in higher income brackets report using internet compared to those in lower income brackets (p for all comparisons <0.0001).

Table 2: Sociodemographic Characteristics of Internet Users versus non-Internet Users

<table>
<thead>
<tr>
<th></th>
<th>Internet User (N=4322), N (%)</th>
<th>Non-Internet User (N=1074), N (%)</th>
<th>Chi Square (df)</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Average Age</strong></td>
<td>54.05 (16.38)a</td>
<td>68.61 (13.7)a</td>
<td>-26.1224 (5243)b</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td><strong>Race/Ethnicity</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-Hispanic White</td>
<td>2648 (87.31)</td>
<td>385 (12.69)</td>
<td>134.5312 (4)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Non-Hispanic Black</td>
<td>501 (74.33)</td>
<td>173 (25.67)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hispanic/Latinx</td>
<td>526 (72.45)</td>
<td>200 (27.55)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asian</td>
<td>191 (85.65)</td>
<td>32 (14.35)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Otherd</td>
<td>174 (83.65)</td>
<td>34 (16.35)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Education Level</strong></td>
<td></td>
<td></td>
<td>816.8747 (3)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>High school degree or less</td>
<td>688 (54.43)</td>
<td>576 (45.57)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Some post-high school</td>
<td>1309 (82.74)</td>
<td>273 (17.26)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>College degree</td>
<td>1278 (91.94)</td>
<td>112 (8.06)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Postgraduate studies</td>
<td>957 (95.32)</td>
<td>47 (4.68)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Income Level (US $)</strong></td>
<td></td>
<td></td>
<td>635.0813 (8)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>0-9999</td>
<td>181 (55.86)</td>
<td>143 (44.14)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10,000-14,999</td>
<td>170 (56.48)</td>
<td>131 (43.52)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15,000-19,999</td>
<td>174 (64.44)</td>
<td>96 (35.56)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20,000-24,999</td>
<td>430 (70.49)</td>
<td>180 (29.51)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>35,000-49,999</td>
<td>306 (81.22)</td>
<td>117 (18.78)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>50,000-74,999</td>
<td>746 (88.49)</td>
<td>97 (11.51)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>75,000-99,999</td>
<td>539 (92.61)</td>
<td>43 (7.39)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>100,000-199,999</td>
<td>853 (96.28)</td>
<td>33 (3.72)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>200,000+</td>
<td>312 (96.59)</td>
<td>11 (3.41)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*aValues are mean (SD) rather than n (%)
*bA 2-tailed t-test was performed
*cA dash indicates that no value was calculated
*dOther includes American Indian/Alaska Native, Pacific Islander, and multiple races
Figure 1 shows how many internet users: watch health-related YouTube videos (37.97%), look up cancer information (21.70%), share health information on social media (14.14%), and participate in online health forums (8.38%).

**Sociodemographic Analysis of Smartphone Users & Purposes of Usage**

Table 3 summarizes the sociodemographic factors of participants who reported using smartphones (N=4172) versus those who reported not using smartphones (N=1121). The average age of smartphone users was 53.4 years (SD 16.2), while the average age of non-users was 69.5 years (SD 13.2) – a significant difference (p<0.0001). More Asians and people who reported “other” races report using smartphones compared to other racial/ethnic populations; more educated people report using smartphones compared to less educated populations; and more people in higher income brackets report using smartphones compared to those in lower income brackets (p for all comparisons <0.0001).

Table 3: Sociodemographic Characteristics of Smartphone Users versus non-Smartphone Users

<table>
<thead>
<tr>
<th></th>
<th>Smartphone User (N=4172), N (%)</th>
<th>Non-Smartphone User (N=1121), N (%)</th>
<th>Chi Square (df)</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Average Age</strong></td>
<td>53.39 (16.15)*</td>
<td>69.54 (13.22)*</td>
<td>-29.9799 (5149)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td><strong>Race/Ethnicity</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-Hispanic White</td>
<td>2450 (81.69)</td>
<td>549 (18.31)</td>
<td>28.0471 (4)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Non-Hispanic Black</td>
<td>499 (75.61)</td>
<td>161 (24.39)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hispanic/Latinx</td>
<td>574 (80.28)</td>
<td>141 (19.72)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asian</td>
<td>199 (90.05)</td>
<td>22 (9.95)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Otherd</td>
<td>175 (85.37)</td>
<td>30 (14.63)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Education Level</strong></td>
<td></td>
<td></td>
<td>456.8328 (3)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>High school degree or less</td>
<td>722 (59.42)</td>
<td>493 (40.58)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Some post-high school education</td>
<td>1235 (79.01)</td>
<td>328 (20.99)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>College degree</td>
<td>1219 (88.59)</td>
<td>157 (11.41)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Postgraduate</td>
<td>910 (91.55)</td>
<td>84 (8.45)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Income Level (US $)</strong></td>
<td></td>
<td></td>
<td>566.1644 (8)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>0-9999</td>
<td>175 (56.09)</td>
<td>137 (43.91)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10,000-14,999</td>
<td>167 (57.59)</td>
<td>123 (42.41)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15,000-19,999</td>
<td>165 (62.26)</td>
<td>100 (37.74)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20,000-34,999</td>
<td>406 (67.55)</td>
<td>195 (32.45)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>35,000-49,999</td>
<td>482 (78.37)</td>
<td>133 (21.63)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>50,000-74,999</td>
<td>739 (88.50)</td>
<td>96 (11.50)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>75,000-99,999</td>
<td>512 (88.58)</td>
<td>66 (11.42)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>100,000-199,999</td>
<td>828 (94.31)</td>
<td>50 (5.69)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>200,000+</td>
<td>310 (97.79)</td>
<td>7 (2.21)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*aValues are mean (SD) rather than n (%); *bA 2-tailed t-test was performed; *cA dash indicates that no value was calculated; *dOther includes American Indian/Alaska Native, Pacific Islander, and multiple races
Figure 2 depicts what proportion of smartphone users (N=4172) use their device to: look up health information for themselves (80.19%); engage with health/wellness apps (54.91%); communicate with a health care professional (50.01%); make an appointment with a health care provider (48.23%); look up medical test results (47.36%); send/receive text messages to/from providers (44.20%); track progress on a health-related goal (43.29%); make decisions regarding treatment plans (42.02%); track health care charges and costs (40.87%); buy medications or vitamins online (33.64%); and look up information regarding the harms of e-cigarettes (9.74%).

Figure 2. Descriptive Statistics Regarding Purposes of Smartphone Usage

**Sociodemographic Analysis of Electronic Wearable Device Users & Purposes of Usage**

Figure 3A shows the proportion of electronic wearable device users (N=1300) who use their device every day (48.91%) and almost every day (20.25%) and Figure 3B shows those who are willing to share device data with providers (80.58%) and family/friends (66.74%).

Figures 3A & 3B. Descriptive Statistics Regarding Purposes of Electronic Wearable Device Usage
Table 4 summarizes the sociodemographic factors of participants who reported using electronic wearable devices (N=1300) versus those who reported not using devices (N=4080). The average age of wearable device users was 49.8 years (SD 16), while the average age of non-users was 59.1 years (SD 16.5) – a significant difference (p<0.0001). More Asians and non-Hispanic whites report using electronic wearable devices compared to less educated populations; more people in higher income brackets report using this HIT compared to those in lower income brackets (p for all comparisons <0.0001).

**Table 4: Sociodemographics of Electronic Wearable Device Users vs non-Electronic Wearable Device Users**

<table>
<thead>
<tr>
<th></th>
<th>Electronic Wearable Device Users (N=1300), N (%)</th>
<th>Non-Electronic Wearable Device Users (N=4080), N (%)</th>
<th>Chi Square (df)</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Average Age</strong></td>
<td>49.79 (15.99)a</td>
<td>59.1 (16.54)a</td>
<td>-17.6480 (5230)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td><strong>Race/Ethnicity</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-Hispanic White</td>
<td>788 (26.01)</td>
<td>2242 (73.99)</td>
<td>11.1655 (4)</td>
<td>0.025</td>
</tr>
<tr>
<td>Non-Hispanic Black</td>
<td>141 (21.01)</td>
<td>530 (78.99)</td>
<td>– c</td>
<td>–</td>
</tr>
<tr>
<td>Hispanic/Latinx</td>
<td>169 (23.28)</td>
<td>557 (76.72)</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Asian</td>
<td>65 (29.15)</td>
<td>158 (70.85)</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Other d</td>
<td>47 (22.60)</td>
<td>161 (77.40)</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td><strong>Education Level</strong></td>
<td></td>
<td></td>
<td>245.8814 (3)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>High school degree or less</td>
<td>130 (10.34)</td>
<td>1127 (89.66)</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Some post-high school</td>
<td>344 (21.77)</td>
<td>1236 (78.23)</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>College degree</td>
<td>447 (32.20)</td>
<td>941 (67.80)</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Postgraduate</td>
<td>349 (34.80)</td>
<td>654 (65.20)</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td><strong>Income Level (US $)</strong></td>
<td></td>
<td></td>
<td>403.6958 (8)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>0-9999</td>
<td>33 (10.19)</td>
<td>291 (89.81)</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>10,000-14,999</td>
<td>18 (5.98)</td>
<td>283 (94.02)</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>15,000-19,999</td>
<td>25 (9.36)</td>
<td>242 (90.64)</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>20,000-34,999</td>
<td>85 (13.96)</td>
<td>524 (86.04)</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>35,000-49,999</td>
<td>122 (19.61)</td>
<td>500 (80.39)</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>50,000-74,999</td>
<td>206 (24.44)</td>
<td>637 (75.56)</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>75,000-99,999</td>
<td>197 (33.85)</td>
<td>385 (66.15)</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>100,000-199,999</td>
<td>346 (39.10)</td>
<td>539 (60.90)</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>200,000+</td>
<td>161 (49.85)</td>
<td>162 (50.15)</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

aValues are mean (SD) rather than n (%); bA 2-tailed t-test was performed; cA dash indicates that no value was calculated; dOther includes American Indian/Alaska Native, Pacific Islander, and multiple races

**Associations Between Purposes of HIT Usage and Sociodemographic Predictors**

Table 5 depicts four separate multivariate logistic regression models analyzing how various purposes of HIT usage (smartphones for health/wellness apps, internet for YouTube videos, smartphones for looking up information, and smartphones for communicating with providers) are associated with sociodemographic characteristics.
Table 5: Purposes of HIT Usage by Sociodemographic Predictors

<table>
<thead>
<tr>
<th>Health/Wellness Apps (N=3458)</th>
<th>YouTube Health Videos (N=4449)</th>
<th>Look Up Health Information (N=3629)</th>
<th>Communicate with Provider (N=3625)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Odds Ratio</td>
<td>P-Value</td>
<td>Odds Ratio</td>
</tr>
<tr>
<td>Age</td>
<td>0.980</td>
<td>&lt;0.0001</td>
<td>0.967</td>
</tr>
<tr>
<td>Race/Ethnicity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-Hispanic White</td>
<td>Ref(^a)</td>
<td>-(^b)</td>
<td>Ref(^a)</td>
</tr>
<tr>
<td>Non-Hispanic Black</td>
<td>1.374</td>
<td>0.005</td>
<td>1.542</td>
</tr>
<tr>
<td>Hispanic/Latinx</td>
<td>1.002</td>
<td>0.981</td>
<td>1.613</td>
</tr>
<tr>
<td>Asian</td>
<td>1.191</td>
<td>0.307</td>
<td>2.471</td>
</tr>
<tr>
<td>Other</td>
<td>1.266</td>
<td>0.200</td>
<td>1.418</td>
</tr>
<tr>
<td>Education Level</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High school degree or less</td>
<td>0.473</td>
<td>&lt;0.0001</td>
<td>0.433</td>
</tr>
<tr>
<td>Some post-high school education</td>
<td>0.711</td>
<td>0.001</td>
<td>0.756</td>
</tr>
<tr>
<td>College degree</td>
<td>0.905</td>
<td>0.332</td>
<td>0.844</td>
</tr>
<tr>
<td>Postgraduate</td>
<td>Ref</td>
<td>-</td>
<td>Ref</td>
</tr>
<tr>
<td>Income Level ($)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0-9999</td>
<td>0.304</td>
<td>&lt;0.0001</td>
<td>0.855</td>
</tr>
<tr>
<td>10,000-14,999</td>
<td>0.311</td>
<td>&lt;0.0001</td>
<td>0.943</td>
</tr>
<tr>
<td>15,000-19,999</td>
<td>0.361</td>
<td>&lt;0.0001</td>
<td>0.990</td>
</tr>
<tr>
<td>20,000-34,999</td>
<td>0.488</td>
<td>&lt;0.0001</td>
<td>0.921</td>
</tr>
<tr>
<td>35,000-49,999</td>
<td>0.487</td>
<td>&lt;0.0001</td>
<td>1.113</td>
</tr>
<tr>
<td>50,000-74,999</td>
<td>0.500</td>
<td>&lt;0.0001</td>
<td>1.027</td>
</tr>
<tr>
<td>75,000-99,999</td>
<td>0.692</td>
<td>0.028</td>
<td>0.854</td>
</tr>
<tr>
<td>100,000-199,999</td>
<td>0.685</td>
<td>0.015</td>
<td>0.991</td>
</tr>
<tr>
<td>200,000+</td>
<td>Ref</td>
<td>-</td>
<td>Ref</td>
</tr>
</tbody>
</table>
\(^a\)Ref=reference group
\(^b\)A dash indicates that no value was calculated

Synthesis of Key Findings

Results show that there was a significant difference in the racial/ethnic composition of internet users: 87.3% of non-Hispanic whites report being users, while only 74.3% of non-Hispanic Blacks and 72.5% of Latinx report being users ($\chi^2=134.53, p<0.0001$). The differences by educational attainment are also striking: only 54.4% of people with a high school degree or less are internet users, while 91.9% of college graduates and 95.3% of post-graduates are users ($\chi^2=816.87, p<0.0001$). As income bracket increases, so does internet usage: only 55.9% of people earning less than $9999 per year are users, whereas 96.6% of people earning $200,000+ are users ($\chi^2=635.08, p<0.0001$).

Next, consider smartphone owners: 90.1% of Asians report owning a smartphone, as do 81.7% of non-Hispanic whites, compared to 75.6% of non-Hispanic Blacks ($\chi^2=28.05, p<0.0001$). Again, as education increases, so does smartphone ownership: only 59.4% of people with a high school degree or less report owning smartphones, compared to 91.6% of people with a postgraduate degree ($\chi^2=456.83, p<0.0001$). A similar trend is seen with income: only 56.1% of people earning less than $9999 per year own a smartphone, while 97.8% of people earning $200,000+ own a smartphone ($\chi^2=566.16, p<0.0001$).
The final HIT type examined, electronic wearable devices, also showed significant differences in racial/ethnic composition of users: Asians (29.2%) and non-Hispanic whites (26%) are the groups with the highest rates of electronic wearable device ownership ($\chi^2=11.17$, $p<0.025$). Once again, the trends by educational attainment and income are extremely clear: 34.8% of postgraduates own such a device, compared to just 10.3% of people with a high school degree or less ($\chi^2=245.88$, $p<0.0001$); and almost half of people earning over $200,000 per year (49.9%) own such a device, compared to just 10.2% of people in the lowest bracket ($\chi^2=403.70$, $p<0.0001$).

In terms of the purposes of internet usage, many participants reported watching YouTube health-related videos (38%), followed by looking up cancer information (21.7%) as a second common purpose. For smartphone users, the vast majority (80.2%) reported using their device to look up health information, and many also reported using it for engaging with health/wellness apps (54.9%), communicating with a provider (50%), making appointments with a health care provider (48.2%), and looking up medical test results (47.4%). Finally, most electronic wearable device users use their device every day (48.9%) or almost every day (20.3%) and are more willing to share device data with providers (80.6%) rather than family and friends (66.7%).

Discussion

As previously mentioned, HIT is a powerful tool to engage not just individuals, but also communities in their own health care. It is important to understand how HIT community engagement strategies can be leveraged to reduce health inequities. The multivariate logistic regression models ran as part of this study can help tailor HIT community engagement strategies to specific sociodemographic groups. Some of the most common purposes of HIT usage (namely, health/wellness apps, YouTube health videos, looking up health information, and communicating with providers) were examined against sociodemographic predictors.

Non-Hispanic Blacks reported a 37% increased odds of using health/wellness apps ($p=0.005$) and a 54% increased odds of watching YouTube health videos ($p<0.0001$) compared to non-Hispanic whites, controlling for age, education, and income. Thus, perhaps HIT engagement strategies that capitalize on apps and videos tailored to this population can help further engage Black/African American communities. Similarly, Hispanic/Latinx populations reported a 61% increased odds of watching YouTube health videos ($p<0.0001$), but a 24% decreased odds of looking up health information ($p=0.039$) and communicating with providers ($p=0.009$) as compared to non-Hispanic whites, controlling for age, education, and income. Thus, for engaging Latinx communities, videos may prove to be a more successful strategy as compared to websites or methods of communication with providers. Finally, note that Asians have a 2.47 times increased odds of watching YouTube health videos ($p<0.0001$) as compared to non-Hispanic whites, controlling for age, education, and income – again suggesting that videos could be a powerful engagement strategy for this community.

One of the most striking findings from this study is the correlation between educational attainment and HIT usage. Specifically, participants with a high school degree or less have a 53% decreased odds of using health/wellness apps ($p<0.0001$), a 57% decreased odds of watching YouTube health videos ($p<0.0001$), a 77% decreased odds of looking up health information ($p<0.0001$), and a 71% decreased odds of communicating with their providers ($p<0.0001$) as compared to postgraduates, controlling for age, race/ethnicity, and income. This makes intuitive sense, as people in this population may not have as much access to these tools and/or knowledge on how to effectively utilize them. This is a very important finding, as it shows that, with status quo HIT tools, a key population of interest – those with low educational attainment – cannot be reached. Significant work must be done to tailor HIT tools to cater to populations with lower health literacy and education rates in order to engage these populations in healthcare, improve health outcomes, and work towards reducing health inequity.

Finally, consider income status and HIT usage. As income bracket decreases, so does HIT usage: those earning $9999 per year or less have a 70% decreased odds of using health/wellness apps ($p<0.0001$), a 47% decreased odds of looking up health information ($p=0.023$), and a 77% decreased odds of communicating with their providers ($p<0.0001$) as compared to those earning $200,000+ per year, controlling for age, race/ethnicity, and educational attainment. This makes sense, given the close relationship between income and educational attainment. Again, these data show that HIT tools must be modified to cater to socioeconomically disadvantaged populations.

Overall, these regression analyses show promising avenues in which HIT community engagement strategies can be tailored for specific populations. By using methodologies that are better suited for certain sociodemographic groups (e.g. more health apps for Black/African American populations, and more health videos for Black, Latinx, and Asian populations), more members will likely participate in HIT usage. Furthermore, this study highlights a need to create more HIT tools inclusive towards populations with low educational attainment and socioeconomic status.
Based on the findings of this investigation, further research must be done to determine if these tailored HIT engagement strategies are indeed effective. The study results will guide future researchers in determining which forms of HIT they should utilize for which sociodemographic groups. It is important to determine if using these tailored strategies does indeed yield better community engagement results, as then these interventions can be implemented at a population-level to work towards achieving health equity.

**Limitations**

It is important to consider some of the limitations of this study. First, the HINTS survey data are cross-sectional, and thus the results of this study indicate correlation rather than causation. Second, the HINTS survey had a relatively low response rate (about 30%) and hence some non-response bias may have been introduced into the data. Finally, there are some limitations to evaluating HIT usage and fully understanding the many ways in which HIT can be used through a population-based survey, which does not go into depth regarding the individual nuances in HIT usage.

**Conclusion**

Overall, this study examines how HIT usage varies by sociodemographic factors through analysis of the HINTS dataset in order to propose tailored technological interventions for various subpopulations. Previous research has shown the value of HIT tools in engaging under-resourced communities in health care and combating health inequities. While prior researchers have designed various HIT interventions and tested them on specific subpopulations, this study will provide guidance for future researchers on which interventions may be best suited for certain sociodemographic groups. Further research must be conducted to determine if these tailored interventions do indeed increase community engagement and can serve as a modality towards achieving health equity.

**References**

DL4Burn: Burn Surgical Candidacy Prediction using Multimodal Deep Learning

Sirisha Rambhatla, PhD¹,∗; Samantha Huang, BS²,∗; Loc Trinh, M.Eng.¹; Mengfei Zhang¹; Boyuan Long¹; Mingtao Dong¹; Vyom Unadkat¹; Haig A. Yenikomshian, MD³; Justin Gillenwater, MD, MS³; Yan Liu, PhD¹;

¹ Computer Science Department, University of Southern California, Los Angeles, CA, U.S.A.
² Keck School of Medicine, University of Southern California, Los Angeles, CA, U.S.A.
³ Southern California Regional Burn Center at LAC+USC, University of Southern California, Los Angeles, CA.

Abstract

Burn wounds are most commonly evaluated through visual inspection to determine surgical candidacy, taking into account burn depth and individualized patient factors. This process, though cost effective, is subjective and varies by provider experience. Deep learning models can assist in burn wound surgical candidacy with predictions based on the wound and patient characteristics. To this end, we present a multimodal deep learning approach and a complementary mobile application – DL4Burn – for predicting burn surgical candidacy, to emulate the multi-factored approach used by clinicians. Specifically, we propose a ResNet50-based multimodal model and validate it using retrospectively obtained patient burn images, demographic, and injury data.

1 Introduction

An estimated 40,000 burns require inpatient admission each year, 30,000 of which are treated at the limited number of burn centers across the country¹. Accurate determination of burn depth is critical for treatment management, in particular surgical decision making, which may impact patient morbidity and mortality². Visual inspection is the most widely used method, which has an accuracy of 64-76% for experienced burn surgeons, a 50% accuracy for inexperienced clinicians, and a tendency of providers to overestimate burn depth³–⁸. Inaccurate burn evaluation poses risks to the patient and the healthcare system. On one hand where overestimation of burn depth leads to unnecessary surgeries and a high rate of inappropriate burn center referrals³,⁸, underestimation leads to delayed surgical management, prolonged healing times, and increased risk for hypertrophic scarring and contractures³,⁸.

Currently there exists a body of work supporting the use of machine learning (ML) in various areas of burn wound diagnosis and management. Few studies have investigated ML models to predict burn wound surgical candidacy. These studies typically use a support vector machine (SVM) model and remain in the investigative stages³,⁶,⁹. Additionally, existing machine learning models predict with a target based only on a wound’s burn depth and associated healing time. A shortcoming of this strategy is the inability to account for the inherent surgical risks as well as patient and injury variables clinicians use in their decision to operate².

We propose a multimodal approach to better emulate the clinical decision making process. We input photographs with patient demographic and injury data to improve the model’s predictive capabilities. We further present our concept for delivering this technology in the form of a mobile application. We seek to translate this technology into a mobile application given the versatility, convenience, and accessibility of hand-held devices as they become more frequently used in the healthcare setting. The overall architecture of the DL4Burn is shown in Fig. 1. The specific contributions of our work are as follows.

• Multimodal Deep learning for Burn image surgical candidacy. We propose a deep learning-based surgical evaluation model which utilizes expert medical opinion along with a deep learning-based model to analyze the images and the multimodal meta-data (such as patient information, first responder and physician’s analysis) to predict the surgical candidacy of a burn wound.

∗Authors contributed equally. Contact corresponding authors Sirisha Rambhatla at sirishar@usc.edu and Samantha Huang at huangsm@usc.edu.
Figure 1: Overview of DL4Burn – a Multimodal Deep learning approach for burn surgical candidacy prediction. Panel (a) shows the multimodal deep learning model specifics. In panel (b) we illustrate the cloud-based workflow of the iOS application being used in the field.

- **Interpretable Assistive Solution.** In addition to providing learning-based recommendations, our method analyzes the outcomes via a state-of-the-art interpretability method to provide insights into the decision-making process of the learning model, and to assist medical evaluations in a trustworthy fashion.

- **DL4Burn iOS Application.** Complementary to the technical contributions, we also introduce the DL4Burn iOS mobile application to be used in the field for burn image and meta-data acquisition, triage, and analysis of images in real-time. This tool will help to accelerate data acquisition and the deep learning model learning, while providing burn surgical candidacy outcomes.

1.1 Related Works

1.1.1 Burn inspection and candidacy evaluation by clinicians

The decision to undergo surgical management is a multi-factorial process often individualized by patient\(^2,10\). Generally, surgeons evaluate the burn depth and percent total body surface area (%TBSA) burned to determine the wound’s healing potential, which is then used to determine if the wound should undergo surgical intervention\(^10,11\). Deep partial to full thickness burns are expected to take greater than three weeks to heal and are therefore surgically managed\(^10,11\). Superficial to superficial partial thickness burns are expected to heal within seven days to two weeks and are non-operatively managed\(^10,11\).

To determine burn depth, clinicians most commonly rely upon visual evaluation, inspecting for color, texture, and capillary refill as well as the presence or absence of sensory loss\(^11,12\). This process is subjective, varied by provider experience, and becomes especially difficult when deciphering mixed partial thickness burns\(^12\). This is especially critical given the differentiation of these burn depths is a baseline for deciding whether or not to operate on a wound\(^12\). As such, there is a need for a more objective method to assist in burn diagnosis and management.

1.1.2 Deep learning for medical imaging

Deep learning has been applied to many areas of medical imaging as an assistive tool for providers. Examples include chest x-ray imaging to detect lung nodules and histologic imaging for identifying high risk breast cancer\(^13\). In burn research, machine learning has been used on various imaging modalities in order to classify burn depth and determine a wound’s surgical candidacy. Specifically, machine learning has been used to classify burn depth from photographic, infrared, and spectroscopic imaging, which is then used to further classify wounds as surgical or non-surgical\(^3,6,9,14–16\). As such, the machine learning models are designed to classify surgical candidacy based on the predicted burn depth. For example, wounds classified as deep partial thickness to full thickness are labeled as surgical wounds while superficial and superficial partial thickness wounds are labeled as non-surgical\(^14,15\).
These studies employ several different machine learning techniques in order to determine burn depth, and by proxy, wounds that would require surgery. Here, Serrano et al. and Yadav et al. used support vector machine (SVM) and feature selection to translate visual cues used by clinicians, like color and texture, into mathematical models.\textsuperscript{9,14} Wang et al. instead, used a ResNet-50 based model to classify wounds by healing time.\textsuperscript{15} Findings from these studies have been promising, with an accuracy ranging from 79%-100% in predicting surgical candidacy, however these have been evaluated on relatively small datasets.\textsuperscript{3,6,9}

It must be noted that a shortcoming of these strategies is the reliance on a simplified view of burn diagnosis – using burn depth and the associated healing times as a proxy for determining burn surgical candidacy. In practice, the decision to operate depends not only on burn depth, but also on individualized patient factors and the inherent risks of surgery.\textsuperscript{2} For instance, patient age influences skin thickness and therefore thermal protection; while pre-existing comorbidities impact the ongoing injury pathophysiology and surgical risks.\textsuperscript{10} Injury characteristics such as the %TBSA burned and presence of sepsis impact thermoregulation, immune responses, and inflammatory cascades, prompting the need to balance aggressive fluid resuscitation with appropriately timed surgical intervention.\textsuperscript{2,10} Additionally, burn surgeons often require multiple evaluations over the initial days of injury to assess for the wound’s evolution and evaluate for areas of healing.\textsuperscript{12} Therefore, a method to evaluate wounds through different time points and a multimodal deep learning model which considers multiple information sources will be a more reliable approach for burn evaluation.\textsuperscript{17}

2 Methodology

We begin by describing our data collection methodology, and then present the proposed multimodal approach.

2.1 Data Collection

We conducted a retrospective chart review at the Southern California Regional Burn Center at LAC+USC on patients admitted between January 2015 and December 2016. Inclusion criteria were patients greater than 18 years at the time of injury, with any form of burn injury, and with photographs taken of their burn wounds during their admission. Available burn wound images without patient identifying data (e.g., eyes, birth marks, tattoos) were uploaded securely to our database. We collected data on patient demographics, injury characteristics, treatment characteristics, and outcomes. The photos were renamed with record number, anatomic location, and post burn day. We recorded these labels in a separate log, where each image was matched with patient injury, treatment, and eventual outcomes data, resulting in a dataset consisting of 174 labelled images requiring (class 1) and 226 images not requiring (class 0) surgical intervention. Note that here the outcome data indicates if the wound eventually required surgery. We treat this outcome as the gold standard. Thus the task is to predict the final outcome (requiring or not requiring surgery) from an image and other characteristics recorded at the time the patient was admitted.

In addition to the images, the features such as age, race (white, black, hispanic, asian), gender (male, female), comorbidities (diabetes mellitus, hypertension, chronic kidney disease, coronary artery disease, congestive heart failure, hyperlipidemia, cerebral vascular accident, liver failure or cirrhosis, psychiatric illness, obesity, chronic obstructive pulmonary disease), anatomic location of burn (arms, hands, thighs, legs, feet, face/head, neck, chest/torso, breasts, back, genitals, buttocks, perineum), mechanism of injury (scald, contact, flame, electrical, chemical, dermatologic), burn depth (superficial, superficial partial thickness, deep partial thickness, full thickness), %TBSA, presence of inhalation injury, presence of wound conversion, and post burn day (PBD), were also collected.

2.2 Preprocessing

We process the data collected to form the multimodal dataset \( D = \{x_i, y_i\}_{i=1}^N \), where \( x_i = \{x_{i}^{\text{Im}}, x_{i}^{\text{Tab}}\} \) denotes the multimodal data samples consisting of images \( x_{i}^{\text{Im}} \) and tabular data \( x_{i}^{\text{Tab}} \), and \( y_i \) denotes the target labels for surgical candidacy, in the following way.

**Image Preprocessing.** Since our retrospective data had inconsistent image sizes, we trained the model using randomly cropped and rotated burn images, which makes the learned model robust to scale and orientation of the burn region. Specifically, we randomly rotate the images \([[-30^\circ,30^\circ]]\), center crop and add random horizontal flips, before we normalize each image, \( x_{i}^{\text{Im}} \) for all \( i \in \{1,\ldots,N\} \). During training, we further randomly crop images in each epoch to
further improve the robustness.

**Tabular Data Preprocessing.** For the tabular data, we encode the categorical data (race, comorbidities, anatomic location of burn, mechanism of injury, burn depth) using one-hot coding. Here, we use a vector of all zeros to indicate a missing entry. We use the binary labels for gender, presence of inhalation injury, presence of wound conversion, and PBD variables. The variables age, burn depth, %TBSA are denoted by their float values, where 0 indicates a missing entry. The resulting features are concatenated to form a vector $x_{i}^{\text{Tab}} \in \mathbb{R}^{62}$, denoting the tabular information.

### 2.3 Multimodal Deep Learning for Burn Wound Evaluation

Multimodal Deep Learning aims to leverage all available information (potentially from varied sources, i.e. multiple modalities) to improve the performance on a machine learning task\textsuperscript{17}. Multimodal information has been effectively used in medical domains such as cervical dysplasia diagnosis\textsuperscript{19,20}, Alzheimer’s disease progression\textsuperscript{21,22} and skin lesion classification\textsuperscript{23}. Adding these multimodal features observed during medical screening improves the prediction accuracy by mimicking the real-world diagnostic procedure, wherein a multitude of factors are considered in decision making. To leverage these, we propose a multimodal deep learning model to predict burn surgical candidacy. Further, we also explain the imaging-based predictions using state-of-the-art interpretability techniques to reveal the decision making process of the deep learning network, critical to assisting clinicians in practice.

The proposed multimodal deep learning architecture for burn evaluation – DL4Burn – shown in Fig. 1 (a), uses two neural networks to arrive at the representations for the image and the tabular data separately. The choice of these neural networks depends on the data they process. For instance, in this paper, we use a deep convolutional neural network (CNN) to process the images and a multi-layer perceptron (with $\tanh(\cdot)$ activation) for the tabular data. We fuse these networks by concatenating the appropriately shaped inner-layer representations, and then pass this through a multi layer perceptron model (with $\tanh(\cdot)$ activation). We obtain the final classification result by using a softmax layer, which outputs the class probabilities\textsuperscript{†}.

### 2.4 Model Specifics and Training

Since the collected burn image dataset is relatively small as compared to the images required to train a deep neural network\textsuperscript{18}, we use ResNet-50 model pre-trained on ImageNet dataset in place of the deep convolutional neural network with three additional trainable multi-layer perceptron layers (with 2048, 1000 and 256 neurons), resulting in a representation of dimension 256\textsuperscript{24}. Next, we learn the representation corresponding to $x_{i}^{\text{Tab}} \in \mathbb{R}^{62}$ using a 2-layer multi layer perceptron each with 32 neurons, resulting in a size 32 representation. We then concatenate these image and tabular representations to form a 288 length vector, and pass it through a 2-layer multi-layer perceptron model containing 128 and 2 neurons respectively, and finally through a softmax layer to calculate the class probabilities. We train the model using stochastic gradient descent optimizer with a learning rate of 0.01 and momentum of 0.9, over 400 epochs with a batch size of 50 using the Cross Entropy Loss function. Overall architecture is shown in Fig. 1 (a).

### 2.5 Interpretability

The decisions made by Neural network-based learning models are fairly opaque and are not interpretable as such. As a result, it may not be clear why a certain prediction was made and what contributed to this specific decision. This is a major issue in healthcare applications, where practitioners need to know which factors contributed to a decision to make an informed choice. To build explainability we leverage a state-of-the-art blackbox model interpretability method in order to build a reliable and robust burn surgical candidacy application\textsuperscript{25}.

### 2.6 Baselines

We compared the performance of DL4Burn with support vector machine (SVM) and neural network-based baselines. Specifically, we use SVM trained on classifier on Histogram of Oriented Gradients (HOG) features\textsuperscript{26} and SVM trained

---

\textsuperscript{†}Class probabilities refer to the probability that a sample belongs to a certain class. In our binary classification setting, this represents the probability that a data sample belongs to class 1.
Table 1: Comparing model performance for burn wound surgical candidacy prediction in terms of area under the curve (AUC) and Accuracy metrics.

<table>
<thead>
<tr>
<th>Method</th>
<th>AUC</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Surgeons</td>
<td>–</td>
<td>0.50–0.76</td>
</tr>
<tr>
<td>HOG-SVM</td>
<td>0.595</td>
<td>0.587</td>
</tr>
<tr>
<td>ORB-SVM</td>
<td>0.548</td>
<td>0.563</td>
</tr>
<tr>
<td>CNN</td>
<td>0.637</td>
<td>0.655</td>
</tr>
<tr>
<td>ResNet-50-based</td>
<td>0.854</td>
<td>0.810</td>
</tr>
<tr>
<td>DL4Burn</td>
<td>0.975</td>
<td>0.938</td>
</tr>
</tbody>
</table>

Table 2: Performance of ResNet-50 and DL4Burn (Neural Network (NN) based methods) across 5-fold cross-validation in terms of area under the curve (AUC) and Accuracy metrics.

<table>
<thead>
<tr>
<th>NN Method</th>
<th>AUC</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Surgeons</td>
<td>–</td>
<td>0.50–0.76</td>
</tr>
<tr>
<td>ResNet-50-based</td>
<td>0.833</td>
<td>0.854</td>
</tr>
<tr>
<td>DL4Burn</td>
<td>0.907</td>
<td>0.975</td>
</tr>
</tbody>
</table>

2.7 Metrics

We report the performance in terms of the area under the curve (AUC) metric, which is a popular metric for evaluating the performance of imbalanced dataset. Specifically, the AUC metric evaluates the area under the Receiver Operating Characteristic (ROC) curve between the True Positive Rate (TPR) and the False Positive Rate (FPR). To construct the curve, we threshold the predicted class probabilities (Section 2.3) at different thresholds between [0, 1], evaluating TPR and FPR at each of these. The AUC metric is especially suited for imbalanced datasets since it measures how good a classifier is at detecting in-class data samples (TPR) while rejecting ones that are out-of-class (FPR). In addition, we also report the accuracy for each of the models. For the best performing models – ResNet-50-based model and DL4Burn (both neural network-based methods) – we also perform and report average and best, AUC and accuracy across 5-fold cross-validation, respectively.

3 Analysis of Results

The AUC and Accuracy results for the proposed and the baseline methods are shown in Table 1. We note that handcrafted features-based techniques perform very poorly on the given dataset. This highlights the need to learn the features from data, which is supported by the results of the neural network-based models. Further, we observe the benefit of using multimodal data for predicting burn surgical candidacy from the performance of the unimodal ResNet-50-based baseline and DL4Burn. Specifically, in Table 2 we show the detailed results between the unimodal (image-only) and multimodal deep learning models. Here, we note the superior performance of the proposed multimodal approach as compared to the unimodal ResNet-50-based baseline. This shows that additional observations made by medical professionals are critical for improving the performance of the deep learning model. It is worth noting that DL4Burn surpasses the human evaluator by a large margin, underscoring applicability of DL4Burn in real-world clinical applications.

Furthermore, we explain the predictions made using the ResNet-50 models for images using Archipelago in Fig. 2. For each image, Archipelago explains how the main effects, and the interactions (the pixels which interact with each other) contribute to the final prediction. Here, the colored patches show if they support (green), are against (red), or are neutral (no color), for a given classification result. These explanations can be leveraged by the physicians to evaluate the surgical candidacy, and aids in the transparency of the learned deep learning model.

4 DL4Burn iOS App

Complementary to the proposed multi-modal deep learning model for burn wound evaluation, we present the prototype of our DL4Burn iOS application, which is an assistive tool designed to aid physicians in burn evaluation and triage. Our DL4Burn iOS app seeks to provide an efficient, cost effective, and reliable adjunct for determining a burn wound’s surgical candidacy. Given the accessibility of mobile phones and their increasingly common use in healthcare settings,
we anticipate its utility in busy emergency rooms, burn unit, and pre-hospital settings. As such, our iOS app has the potential to reduce healthcare errors, their associated costs, as well as streamline the delivery of burn care.

4.1 Functionalities

The functionalities of the app consist of a physician-based interface that enables them to upload burn images and patient demographic and injury data in order to track and manage burn patients. This is a cloud based app, in which uploaded data will be stored in the Google Firebase cloud (Fig. 1) or other cloud-based services. Here, we plan to use the Google Firebase, for its secured cloud-based storage service for mobile application development and deployment. Specifically, for HIPPA compliance and security, we utilize its built-in security authentication and authorization‡.

To begin, the main screen of the app prompts physicians to sign up or log in with their credentials (Fig. 3(a)). Once they are logged in, the app takes the user to a portal in which they add or select from existing patient profiles. Here they can either view the patient’s existing data or upload new data (Fig. 3(b)). To capture a new image for a patient, physicians can select ‘Camera’ on the navigation bar, where they will be able to take a photo and enter the associated patient demographic and injury data (Fig. 3(c)). A random string is generated to uniquely identify each image. This same randomly generated string is used as a key to store details about the patient as well in the real-time database. Screenshots for Google storage and Google Real-time database are given in Fig. 4. Once the image and data are uploaded, the information is stored on the cloud database and immediately fetched into the app. The physician can toggle to the ‘Users’ tab to view the uploaded image at any time (Fig. 3(d)).

We anticipate physicians will be able to capture burn wound images and injury data throughout a patient’s treatment period in order to more closely monitor the wound while also further training the deep learning model. As such, our iOS app enables physicians to keep an organized and photographic record of a wound’s progression and store this data for future evaluations in the treatment process. Simultaneously, these data points will further boast the predictive capacities and inform future applications of our deep learning model.

‡See https://firebase.google.com/docs/storage/security for details.
Figure 3: Overview of DL4Burn landing screens and Camera tool in the iOS app. Panel (a)-(d) show the registration/log in page, and the main landing screen of the app, camera functionality to capture image and enter patient details, and the main landing screen of the app with the image just uploaded, respectively.

5 Discussion

Summary. Determining whether a burn wound will require surgery is an important step in the management of burn patients. Appropriate timing of surgery is critical for optimizing patient outcomes, minimizing the risk for hypertrophic scarring and contractures, and minimizing healing time and hospital length of stay. Conversely, correctly identifying surgical burn wounds spares patients from unnecessary surgeries, their associated risks and healthcare costs.

Incumbent on the surgical decision-making process is an accurate burn wound evaluation. Clinicians most commonly utilize visual inspection to determine burn depth, using it as an estimator for healing time. Notably, the accuracy of this method ranges from 50-76% and is dependent on clinician experience. Burn depth and the associated healing time is further used to determine treatment strategy. Superficial to superficial partial thickness burns are considered nonsurgical wounds and deep partial thickness to full thickness burns are considered surgical wounds. Existing machine learning literature takes advantage of this classification method, designing models that determine burn depth and in turn surgical candidacy. These models have achieved accuracies of greater than 80%, demonstrating the potential role for machine learning as an adjunct for burn diagnosis.

However despite this success, the decision to operate is less opaque than simply determining burn depth. Clinicians must also consider the surgical risks, patient factors, and associated injury characteristics to decide whether to operate. Therefore our proposed multimodal approach addresses these limitations by leveraging patient demographic and injury data with photographic images, a process that assimilates more to that of clinicians. Multimodal deep learning involves the integration of data from more than one domain, used to improve classification by integrating information from each data source. In this study, we demonstrated that a multimodal approach can effectively combine the available information and physician input to provide superior performance as compared to a single domain approach.

Limitations and Future Work. As with any study, it is important to consider the limitations. The results presented in this work were based on a relatively small dataset, which may have led to slight overfitting. To counter this, we plan to leverage the proposed DL4Burn mobile application to scale data collection and hence improve model generalization. Next, due to the retrospective nature of our study, we understand that our data is dependent upon the accurate and complete documentation in the records. As such, the current multimodal features formed in our analysis were drawn from the evaluations of medical professionals available in the charts. Our future work will focus on extracting these features from prospectively obtained data and observations entered through the application. Next, in this work we utilize the final outcome, i.e. if the wound eventually required surgery or not. Specifically, our procedure relies on
Figure 4: Overview of data stored in Google Firebase. Panel (a) and (b) show the image stored in Google Storage and the data about patient along with the image description and url of the image stored in real-time database, respectively.

these final assessment as opposed to initial assessments, since initial assessments are more likely to be incorrect. In our future studies, we plan to record initial assessments to analyze how clinicians’ recommendations change over time. Furthermore, it would also be interesting to equip the learning algorithm with specific domain knowledge used by clinicians to evaluate and improve surgical candidacy.

Another line of work can also study the counterfactuals and effect of interventions. To enable such analysis, we plan to add functionality to track wound progression and interventions over time. This feature can also be used to assist remote assessment of wounds and monitoring surgical candidacy on-the-fly.

Finally, the COVID-19 pandemic has also highlighted the need to develop tele-medicine capabilities for burn wound monitoring. Motivated from this, we are also planning to develop an interface for patients wherein medical professionals can track wound progression remotely. We also seek to analyze successive images from patients to track wound healing, progression, and scarring outcomes so that we may improve and inform future iterations of the software.

Conclusions. Burn wound surgical candidacy involves complex decision making which needs to be customized for each patient, making it particularly challenging to standardize. In a number of cases, the decision to operate (or not operate) may result in long-term medical complications. As a result, there is a need to accurately predict surgical candidacy based on historical patient records to improve patient outcomes. The proposed DL4Burn model and mobile application leverage the expert analysis along with burn images to build reliable multimodal deep learning-based assistive tool for surgeons to improve decision making and patient care.

6 Acknowledgments

The authors would like to graciously acknowledge Joshua Lin, Megha K. Sheth, and Justin Dang for their contribution towards data collection and annotation. The analysis was conducted after receiving approval from our institutional review boards (IRB) number HS-21-00201.

References


Development of an Online Contraceptive Decision Aid for College Women

Molly Redman; Jenny Brian, PhD; Dongwen Wang, PhD
Arizona State University, Tempe, Arizona

Abstract

Lack of knowledge in highly efficient contraceptive methods led to low rates of adoption and misuse of these methods by young women. The existing online tools for contraceptive decisions have flaws. To address this critical need, we developed a prototype online contraception decision aid for college women. For this purpose, we conducted a focus group interview for needs assessment. We designed a scoring system to provide accurate and customized recommendations based on a user's preferences. We implemented the tool with specific functions to collect users' needs and preferences in selecting specific contraceptive methods, to present the customized recommendations, to provide side-by-side comparison of all contraceptive methods, and to recommend additional resources. Preliminary data seem to indicate positive evaluations of the tool. Future work is required to examine the generalizability of the findings and to have full implementations of the tool for real world use.

Introduction

Background

Birth control, commonly referred to as contraception, is defined as any medication or device to prevent pregnancy. There are a variety of contraceptive methods, such as condoms, birth control pills, and intrauterine devices (IUDs), which can be administered in different fashions. Some methods are highly effective at preventing pregnancy, while others are not due to their nature or high probability of user error. Many methods can be used simultaneously with condoms as extra protection against pregnancy, HIV, and other sexually transmitted infections (STIs). Beyond pregnancy prevention, contraceptive methods are beneficial in other aspects of women's health, including regulation of hormones, management of periods, and control of hormonal acne.

Each birth control method is different. Even the same method may work differently for each individual woman. Previous studies have shown that lack of knowledge and awareness of highly efficient contraceptive methods results in their low rates of adoption and misuse, especially in young women. Despite having a higher fail rate compared to other methods, the birth control pill is the most commonly prescribed method among college women. Approximately 77% of this population have reported using birth control pills as their primary choice for contraception. Yet there are many other available methods.

There are two popular online contraceptive decision aids – the Planned Parenthood Birth Control Quiz and the Bedsider Comparison Matrix. The Planned Parenthood tool consists of a short quiz that evaluates a user’s preferences, lifestyle, and opinions on selected methods. At the conclusion of the quiz, the tool recommends three birth control methods that account for the user’s inputs. There are also videos that serve as additional educational resources for those interested in learning more about the recommendations. The Bedsider Comparison Matrix is a large grid that allows a user to compare and contrast many different contraceptive methods at the same time, based on criteria such as “prevents pregnancy” and “party ready.” There is another feature called “Build Your Own” that allows the user to compare up to three different methods at one time. This feature enables the user to learn more about methods that they are specifically interested in.

Preliminary Study on Accuracy and Usability of the Existing Online Tools for Contraceptive Decision

We conducted a preliminary study to assess the accuracy and usability of the tools provided by Planned Parenthood and Bedsider. Here accuracy reflects how well the tool’s recommendation complies with what is expected, i.e., an accurate recommendation would suggest a method that fits what a user has identified as the most important needs and preferences. Meanwhile, usability refers to how easily the tool allows for a user to receive the recommendation, including components such as navigation, layout, and general graphic design. The research aim of this preliminary study was to evaluate the two existing online tools for contraceptive decision, identify any gaps in accuracy (only applied to the Planned Parenthood tool) and usability, and if so, propose potential improvements.

For this purpose, we conducted a literature review on birth control usage and usability heuristics. We then performed a usability walkthrough for both tools to examine whether they were user friendly. Lastly, we developed a set of test
cases that enumerated the possible combinations of decision factors such as to provide a quantitative evaluation on the accuracy of the Planned Parenthood quiz.

We found that the language used for the Bedsider matrix was understandable by both experienced and inexperienced users. Furthermore, there was consistent use of symbols across the platform, therefore allowing for easy transition between pages. However, the tool heavily relied on these symbols and did not provide labels to clearly indicate their meaning. As a result, there was a heavy reliance on recall rather than recognition, a flaw in usability. The information provided in the Bedsider tool was clear. However, there was no opportunity to create a customized result based on user needs, which was another limitation.

In contrast, the Planned Parenthood quiz had a focus on providing customized recommendations to a user. There were many additional information sections throughout the quiz to help the user understand the rationale of certain questions. However, as the user progressing through the quiz, there was no opportunity to go back to correct a mistake. Furthermore, the recommendations at the end did not always accurately reflect the needs or preferences of the user.

To assess the accuracy of the Planned Parenthood quiz, we developed a set of 200 test cases with all possible combinations of the decision factors. We applied the quiz to these 200 test cases, manually reviewed the recommendations generated for each, and examined: (1) the adherence to CDC guidelines; and (2) the accurate reflection of user’s preferences. Of the 200 test cases, 174 (87%) adhered to the CDC guidelines. Among the 177 test cases with an expression of strong user preferences, only 108 (61%) of the recommendations accurately reflected them.

These results have revealed that: (1) the Planned Parenthood quiz and the Bedsider’s matrix each has its own usability flaws; and (2) there is a gap in accuracy of the recommendations generated by the Planned Parenthood tool in fitting an individual’s preferences.

Study Objectives

To address the issues identified from the above preliminary research, we decided to launch a study to develop a new birth control decision aid, My Contraceptive Choice (MCC), with the following aims: (1) generating accurate recommendations that comply with CDC guidelines and individual user’s preferences; (2) presenting clear and concise information; and (3) providing user control in navigation and the opportunity to correct mistakes. We selected to focus on college women as the target user population of the MCC tool. Here college women are defined as females between the age of 18 and 24 years old and currently attending a four-year public university. This user population has major needs in contraception and yet, as discussed earlier, presented knowledge gaps in selecting the appropriate methods. We report in this paper the development process for the MCC tool. The evaluation of the MCC tool is ongoing. We will report the details of the evaluation study and the results in the future.

Methods

The study started with a focus group interview for the target users to solicit their inputs on the factors they considered when choosing a contraceptive method as well as the desired functions for a new tool to assist their decision making. These inputs were then used to guide the development of the MCC tool.

Due to the COVID-19 pandemic, we decided to conduct the focus group interview online through Zoom. For this purpose, we sought the potential participants for the focus group through the student email lists at the Arizona State University (ASU), obtained consent, and recruited ten female students who were between 18 and 24 years old. Each participant received a $35 Amazon gift card as a compensation for the two-hour focus group session. The ASU IRB approved this study.

We prepared a series of questions for the focus group participants, including: (1) the types of birth control methods they used regularly; (2) selection of specific methods; (3) experiences of using specific methods; and (4) feedback on the two existing online tools for contraceptive method selection (Bedsider and Planned Parenthood). We recorded the focus group session through Zoom. For data analysis, we reviewed the transcript for sentiments that were heavily agreed on with regard to specific questions. We noted the phrases or ideas that were repeated multiple times. We paid special attention to the factors that impacted the participants’ decision making as well as their opinions (either positive or negative) on specific features of the two existing tools.

The design of the new online tool MCC was guided by the CDC guidelines and other related medical literature, the results from the preliminary study on the two existing tools, and the findings from the focus group. Specifically, we envisioned that the MCC tool should provide three important functions: (1) customized recommendations based
on a user’s needs and preferences; (2) side-by-side comparison of multiple contraceptive methods such as to facilitate a user’s decision making; and (3) links to additional resources for further education on specific contraceptive methods, purchase of specific devices, or clinics for medical consultations. For the function described in (1), we relied on the CDC guidelines, the other related literature, and the findings from the focus group interview to select/organize the specific factors and to design an algorithm for customized decision aid. For the function described in (2), we reused certain features from the Bedsider tool deemed to be user friendly, and meanwhile added or revised other features such as to address the usability issues identified from the preliminary study and the focus group. For the function described in (3), we reorganized the existing resources listed on the Bedsider and Planned Parenthood tools, and included additional resources as suggested by the focus group participants.

The initial implementation of the MCC tool was for proof of concept. We thus focused primarily on the user side. For this purpose, we designed the MCC as a client-side web application, based on Hypertext Markup Language (HTML), Cascading Style Sheets (CSS), and JavaScript functions. We present the details on implementation of the MCC tool in the next section.

**Results**

*Findings from the Focus Group Interview*

The focus group included participants with various experiences in use of contraception methods, ranging from inexperienced to users of multiple contraceptive methods including condoms, IUDs, and birth control pills. The motivations behind using these specific methods were primarily for pregnancy prevention, regulation of hormones, and/or management of menstruation and its side effects. Some participants used the dual method (condom + either IUD or birth control pill) for extra protection, while others selected to use only one method at a time. Furthermore, many participants agreed that they felt comfortable using their respective contraceptive methods because their close family members or friends had positive experiences using the same methods.

The participants indicated that they wished they were aware of certain side effects and complications prior to starting a contraceptive method. They agreed that much of the information and resources they received previously were hard to understand and not very helpful. A common side effect participants reported experiencing with contraceptive methods, especially the birth control pill, was weight gain or the difficulty to lose weight. Other medical conditions participants reported as side effects from their contraceptive methods included high blood pressure/hypertension, anxiety, and depression. Although these side effects were hard to manage, some participants said they would still use the same method but wished to be more educated.

However, not all side effects discussed were negative. Many participants agreed that being on the pill reduced their signs and symptoms of acne, a frequently seen medical condition in this population. Additionally, other participants stated that their contraceptive methods helped manage painful menstruation side effects and some even reported having their periods stop altogether. Some participants expressed that they liked the latter side effect, while others did not want this result. Additionally, two participants used their birth control methods to maintain health conditions such as polycystic ovary syndrome (PCOS) or endometriosis.

After taking the Planned Parenthood Birth Control Quiz, the majority of the participants expressed dissatisfaction in their recommendations. From the accuracy perspective, the recommendations did not seem to reflect their inputs on specific needs or preferences expressed during the quiz. The star rating system used by the tool to indicate how well a method fitting for an individual seemed to be somewhat cryptic, as many participants received the maximum number of stars for recommendations of multiple methods but were unsure which fit them better. In regard to the questions asked in the quiz portion, participants were not sure why some questions were needed as they did not seem to have a clear impact to the final recommendation. Including these questions made the quiz long and repetitive, and therefore difficult to remain engaged in.

The participants indicated that the Bedsider matrix was confusing and difficult to decipher at the first glance. They agreed that there was a steep learning curve to understand the purpose of the tool and to use it efficiently. However, once this phase passed, the information was very useful and easy to understand. The “Build Your Own” feature of the matrix was also confusing to many and did not seem that helpful when attempting to compare different methods. One of the largest usability issues was the difficulty in remembering the meaning of specific symbols, which we noted in the preliminary study.

When discussing the development of MCC, participants strongly agreed that they would like to see a mixture of Planned Parenthood’s quiz and Bedsider’s matrix. They wished for a decision aid tool that utilized their inputs to
create customized recommendations, but also wanted the ability to compare and contrast to other methods on a general scale. They also wished for features that clearly demonstrated how their inputs were used to make the recommendations and how the methods fit them as an individual.

Implementation of MCC – Gathering Data of User Needs and Preferences

Based on the CDC guideline, additional medical literatures, the preliminary study on the two existing online tools, and the findings from the focus group, we decided to organize the three main functions of MCC in a sequential order, i.e., recommendations of contraceptive methods with customization to user needs, side-by-side comparison of all contraceptive methods, and additional resources. To use the MCC tool, one can start at the homepage with a brief introduction on the purpose of the tool. From there the tool can be launched.

The MCC tool focuses on a list of contraceptive methods relevant to the target user population of college women, including condom, the copper IUD, the hormonal IUD, the ring, the mini pill, the combination pill, the implant, the patch, the shot, and fertility awareness method (FAM)². The customized recommendations are based on users’ preferences in selecting contraceptive methods, prior experiences in using specific ones, users’ medical history with potential interferences with certain methods, and additional factors that may impact the selection. To solicit the user inputs from these aspects, we have designed four distinct pages - Personal Preferences, User Experiences, Medical History, and Additional Factors, each with a set of questions.

On the Personal Preferences page, we use four Likert Scales (very unimportant, unimportant, neutral, important, and very important) to determine a user’s priorities on specific features of a contraceptive method. These features include cost effectiveness, preventing pregnancy, managing periods and side effects, and (low) possibility of weight gain. A screenshot of the Personal Preferences page is shown in Figure 1.

![Personal Preferences Page](image1.png)

**Figure 1.** A screenshot of the Preferences Page of MCC.

On the User Experiences page, users have the opportunity to indicate the specific methods they have used before. By selecting a specific method used before, users can further indicate the past experience (negative, neutral, or positive) and whether they are considering using it again in the future. A screenshot of the User Experiences page is shown in Figure 2.

On the Medical History page, users can express concerns on a list of their medical conditions (acne and breakouts, blood clotting disorders, depression or anxiety, hypertension or high blood pressure, polycystic ovary syndrome (PCOS) or endometriosis, and treatment of sexually transmitted infections (STIs)) that may interfere with use of specific contraceptive methods. A screenshot of the Medical History page is shown in Figure 3.

On the Additional Factors page, there are three additional questions: (1) how comfortable the user feels with the insertion of a foreign body in vagina (very uncomfortable, uncomfortable, neutral, comfortable, or very comfortable); (2) the level of hormones that the user feels most comfortable using (no hormones, one hormone, two hormones, or don’t know/no preference); and (3) how often the user wants to maintain the birth control methods (daily, weekly, monthly, yearly, or don’t know/no preference). A screenshot of the Additional Factors page is shown in Figure 4.
Figure 2. A screenshot of the User Experiences page of MCC.

Figure 3. A screenshot of the Medical History page of MCC.

Figure 4. A screenshot of the Additional Factors page of MCC.
Scoring System to Provide Customized Recommendations for the Top Ranked Contraceptive Methods

To provide customized recommendations for the contraceptive methods that fit with a specific user’s needs or preferences, we have developed a numerical scoring system. This scoring system assigns scores to each contraceptive method based on a user’s answers to the questions on the above four pages. As the user moves from one question to the next, the scores for each contraceptive method are added up. After completion of the data collection on user needs and preferences, the three contraceptive methods with the top scores are the recommendations to the user. We discuss the details of this scoring system next.

On the Personal Preferences page, each question focuses on a specific feature for decision, i.e., cost effectiveness, preventing pregnancy, managing periods and side effects, and (low) possibility of weight gain. Only the contraceptive methods in compliance with that feature are scored. Here, the methods that are considered cost-effective include condom, FAM, mini pill, combination pill, the patch, and the ring. Methods that are good at preventing pregnancy include condom, mini pill, combination pill, the ring, the implant, copper IUD, and hormonal IUD. Methods that help reduce periods and their side effects include mini pill, combination pill, the ring, the implant, and hormonal IUD. Methods that have a low possibility of weight gain include condom, FAM, and copper IUD. For a specific question on a particular feature, a relevant method is awarded a point of -2, -1, 0, +1, or +2, corresponding to the user’s answer as very unimportant, unimportant, neutral, important, or very important, respectively. For example, if a user’s response to the question on low possibility of weight gain is important, the three relevant methods, condom, FAM, and copper IUD each is awarded +1 point.

On the User Experiences page, if a user has a positive experience for a specific contraceptive method, that method is awarded +8 points, which is equivalent to the maximum positive points of all their preferences combined from the previous page. In other words, a positive experience for a specific method overrides all the previous opinions against it. An additional +2 points are rewarded if the user indicates to use the method again. If a user has a negative experience for a specific method, that method is awarded -8 points, i.e., overriding all the previous preferences for it. If a user has no interest in using a method again, it will be removed from the final recommendation list altogether.

On the Medical History page, if a user has checked acne and breakouts or polycystic ovary syndrome (PCOS) or endometriosis, each contraceptive method that can help address these concerns, including the combination pill, the patch, and the ring, is awarded +2 points. The other medical conditions may lead to potential contraindications when selecting specific methods. Therefore, if a user has expressed concerns on one of those medical conditions, all its contraindicating birth control methods will be removed from the final recommendation list. Specifically, for blood clotting disorders, depression or anxiety, and hypertension or high blood pressure, the contraindicating methods include the combination pill, the patch, and the ring. For treatment of sexually transmitted diseases (STIs), the contraindicating methods include hormonal IUD and copper IUD.

On the Additional Factors page, the scoring for the first question (how comfortable with the insertion of a foreign body in the vagina) is similar to the Likert Scale system used for the questions on the Personal Preferences page. Specifically, a relevant method (hormonal IUD, copper IUD, and the ring) is awarded a point of -2, -1, 0, +1, or +2, if the answer to the question is very uncomfortable, uncomfortable, neutral, comfortable, or very comfortable, respectively. For the scoring of the second question (level of hormones), each of the relevant method for a specific answer (no hormone, one hormone, or two hormones) are awarded +2 points. The relevant methods for no hormone are condom, FAM, and copper IUD; the relevant method for one hormone are the implant, the mini pill, the shot, and the hormonal IUD; and the relevant methods for two hormones are the patch, the combination pill, and the ring. For scoring of the final question (frequency to maintain a method), each of the relevant method for a specific answer (daily, weekly, monthly, or yearly) is awarded +2 points. Here the methods that must be maintained every day are condom, FAM, the mini pill, and combination pill; the methods that must be maintained on a weekly basis are the patch and the ring; the method that is maintained on a monthly schedule is the shot; and the methods that must be maintained on a multi-year schedule are the implant, the copper IUD, and the hormonal IUD.

After the scorings for each of the questions on the four pages are completed, every contraceptive method now has a total score. The MCC tool generates a list of recommendations based on the top three methods with the highest scores that match with a user’s preferences, prior experiences, medical conditions, and other needs. The complete algorithm is illustrated in Figure 5.
Figure 5. The algorithm of the scoring system to make customized recommendations.
Graphical Presentation of the Recommended Contraceptive Methods and Side-by-Side Comparison of All Methods

The feedback from the focus group interview indicated that the users preferred a tool with both the customized recommendations and the side-by-side comparisons of all contraceptive methods. When designing the function to present the customized recommendations, we decided to merge it with the side-by-side comparison of all methods to achieve efficiency. For this purpose, we divided the customized recommendations into two segments (pages), i.e., a graphical presentation and a detailed description in text.

Immediately after users have completed the quiz (the first four pages), they are directed to the first segment of the presentation, which uses a tabular design similar to Bedsider’s matrix. The columns of this table include all the contraceptive methods, while the rows list the major factors considered by young women when selecting a birth control method. Each cell of this table is filled with either a green checkmark (yes), a red cross-mark (no), or a gray circle (sometimes/maybe), indicating how well the corresponding method (column) matches with the factor (row). In the same table, the columns corresponding to the top three recommended contraceptive methods are highlighted in blue color. Meanwhile, the rows corresponding to the factors deemed as important to a specific user (based on the user’s inputs from the previous four pages) are bolded. The user can also hover the mouse over a specific cell to learn more about the corresponding contraceptive method from the tooltip. A screenshot of this page is shown in Figure 6.

**Figure 6.** The customized recommendations are combined with the side-by-side comparison of different methods.

Details of the Recommended Contraceptive Methods and Additional Resources

The second segment of the results section displays the details of the customized recommendations for a user. For this purpose, we organize the top three recommended methods in separate columns, each containing the detailed information pertaining to what the method is, how it fits the user, how to use the method, as well as resources for further education and finding a pharmacy/clinic. In the section to explain how a method fits the user, the system dynamically generates bullet points based on the user’s inputs. For example, if a user wants a method that is very good at preventing pregnancy, a bullet point will show up here indicating how effective the selected method can address this need. A screenshot of this segment is shown in Figure 7.

The last function of MCC that we envisioned was to provide additional resources for all contraceptive methods. Since this part is similar to the last section of the recommendations described above, we decided to merge the function of Additional Resources with the previous segment as an appendix. This appendix includes all contraceptive methods, each with two sections of additional resources: (1) the Information section links to the Planned Parenthood and Bedsider websites for more educational resources; and (2) the Find a Pharmacy or Clinic section links to sites to locate a pharmacy/clinic or to make a purchase online.

Evaluation of MCC

The evaluation of the MCC tool is ongoing. The preliminary results from an online survey of 150 users indicated that: (1) 112 (75%) believed that the recommendations they received were appropriate and reflected their personal preferences; and (2) 136 (91%) thought that the layout of MCC made it easy to navigate the tool. We will report the details of the evaluation study and the complete results in a future paper.
Discussion

In design of the MCC tool, we have obtained important insights from the focus group participants with regard to why they use specific contraceptive methods. Specifically, we have found that the most important factors considered by a college woman when selecting a contraceptive method are cost effectiveness, pregnancy prevention, management of periods and their side effects, and little weight change. Neither Planned Parenthood nor Bedsider has included the last factor of little weight change in its tool, but it is indeed an important consideration by many college women.

The focus group members unanimously agreed that MCC should combine the best aspects of Planned Parenthood’s quiz (customized recommendations) and Bedsider’s matrix (side-by-side comparison). Our design directly reflected this request. In collection of users’ needs and preferences, we included additional factors, such as the potential concerning medical conditions, which were not in the Planned Parenthood quiz. On the Recommendation Page, we emulated Bedsider’s matrix but made significant revisions, including: (1) removed factors that typically were not applied to the targeting user population, such as can be used while breastfeeding; (2) removed methods that seemed outdated or inefficient, such as the diaphragm and withdrawal; (3) highlighted the top three recommendations based on a user’s needs and preferences; and (4) highlighted the factors deemed to be important for a specific user. We therefore achieved the goal to combine the customized recommendation and side-by-side comparison to show the pros and cons of each method with an efficient design.

A primary focus of the MCC design was to ensure that: (1) the user had the opportunity to indicate their previous experiences with birth control methods, and (2) the tool’s recommendations of specific methods should reflect the user’s experiences. While the Planned Parenthood quiz was supposed to have a similar feature, the focus group feedback and test case data both suggested that the feature did not work. We therefore designed an algorithm that combined the numerical scoring and hard rules for this purpose. If a user had good experiences using a specific contraceptive method, the system would award +8 points to that method. On the other hand, if a user had bad experiences using a method and would not consider using it again, the system would exclude that method from the final recommendations. It seems this strategy is able to provide customized recommendations, as indicated by the preliminary data from the online survey. Detailed analysis is needed here, and we expect to report the complete results in a future paper.

We have made the navigation of the MCC tool easy by using the “continue” and “back” buttons on each page. A user can go back to a previous page to review and correct any mistakes in data entry. The preliminary data seemed to indicate that users had very positive feeling on navigation.

There are a few limitations in this study. First, the focus group participants were recruited from ASU’s College of Health Solutions and Barrett Honors College. Many students in these two colleges study in health-related disciplines, and thus may have more knowledge on birth control. In addition, the focus group approach has its intrinsic limitations, such as small sample size and participant’s reluctance to express opinions different from the majority. Generalization of the findings, such as the preferred structure and language used for the tool, from this focus group to other populations should be further studied. Second, in development of the algorithm, we did not consider insurance coverage when assigning the cost-effective methods. When designating the methods with low possibility of weight...
gain, we took a conservative approach to exclude the methods with controversial evidence\(^\text{15,16}\). The scoring system, in particular, the weight for a specific factor, was based on heuristics. The robustness of the recommendations generated by the MCC tool therefore should be further tested. In the ongoing evaluation study, we have included quantitative measures on performance of the system based on simulated test cases. We will report the results in a future study. Third, the current implementation of the MCC tool is for proof of concept, focusing only on the client side. Full implementation of the system is needed to capture the actual use of the tool in real world. This could be another direction for future work.

**Conclusion**

We have successfully developed a prototype of the MCC tool to help college women select birth control methods. We have leveraged the inputs from a focus group, integrated certain useful features from two existing tools, and built an efficient, easy-to-navigate system. We have designed a scoring system for customized recommendations of contraceptive methods that are based on users’ needs and preferences. Preliminary data from an ongoing study seem to indicate positive evaluations of the tool. Future work is required to examine the generalizability of the findings to other populations and to have full implementations of the tool for real world use.

**Acknowledgements**

The Barrett Honors College at ASU provided funding support for this honors thesis research by the first author. We thank the students who participated in the focus group for their contributions to this study.

**References**

Comparing Scribed and Non-scribed Outpatient Progress Notes

Adam Rule, PhD1, Sarah T. Florig, MS1, Steven Bedrick, PhD1, Vishnu Mohan, MD1,
Jeffrey A. Gold, MD1, Michelle R. Hribar, PhD1
1Oregon Health & Science University, Portland, OR

Abstract

Working with scribes can reduce provider documentation time, but few studies have examined how scribes affect clinical notes. In this retrospective cross-sectional study, we examine over 50,000 outpatient progress notes written with and without scribe assistance by 70 providers across 27 specialties in 2017-2018. We find scribed notes were consistently longer than those written without scribe assistance, with most additional text coming from note templates. Scribed notes were also more likely to contain certain templated lists, such as the patient’s medications or past medical history. However, there was significant variation in how working with scribes affected a provider’s mix of typed, templated, and copied note text, suggesting providers adapt their documentation workflows to varying degrees when working with scribes. These results suggest working with scribes may contribute to note bloat, but that providers’ individual documentation workflows, including their note templates, may have a large impact on scribed note contents.

Introduction

In the decade since the HITECH Act catalyzed electronic health record (EHR) adoption in the United States,1 there has been growing concern that providers spend too much time documenting patient care.2 The typical primary care physician now spends as much time in front of their EHR as with patients,3,4 and there is growing evidence that more EHR time, especially after work, contributes to higher rates of physician burnout.5,6,7 There is an urgent need to reduce EHR burden,8 which has led many physicians to employ medical scribes—clerical staff who help document patient care.9 Scribes are often primarily employed to document exam room conversations in clinical notes, but scribes may also perform a wider range of tasks based on their clinical training or the provider with whom they work.10 Many scribes are students with little prior medical experience, though some have prior training as nurses, technicians, or medical assistants and thus have license to perform some clinical as well as clerical tasks.

Prior work has found that working with scribes often reduces provider documentation time while increasing physician satisfaction and clinic volumes.11,12 Yet, few studies have examined how working with scribes affects clinical notes themselves. One study found provider’s perceptions of chart quality and accuracy increased when working with scribes,13 while another found scribed notes in primary care clinics were of slightly higher quality when measured on the PDQI-9 scale.14 However, other work found too much variability between raters to use the PDQI-9 to evaluate scribed note quality in emergency medicine.15 Likewise, one simulation study found significant variation between scribes in the completeness, quality, and accuracy of their notes.16 More evidence is needed to assess how employing scribes affects the contents of clinical notes. In particular, there is a need to assess the variation of scribe impact across providers and specialties to help illuminate best practices for collaboratively documenting patient care. Most prior work has instead examined scribes working with a handful of providers in a single specialty.10

This retrospective cross-sectional study aims to help fill this knowledge gap by examining how scribed notes differ from those written without scribe assistance. Since scribes often focus on exam room interactions, we hypothesize:

1. H1: Scribed notes are longer than notes written without scribe assistance
2. H2: Scribed notes contain more manually typed text, but similar amounts of copied and templated text

To test these hypotheses, this study examines over 50,000 notes written by 70 providers in 27 specialties across 2017-2018, both with and without scribe assistance. We find scribed notes are significantly longer, containing more manually typed text, but also more templated and copied text. We also find substantial variation between providers in how working with scribes affected their mix of typed, templated, and copied note text. Together, these findings suggest working with scribes may contribute to note bloat, but that providers’ individual documentation workflows and note templates may have a large impact on the contents—and thus the quality and accuracy—of scribed notes.
Methods

This research was conducted at Oregon Health & Science University, a large academic medical center in Portland, Oregon. OHSU implemented a commercial EHR for billing, documentation, and practice management starting in 2005 (Epic Systems, Verona WI) and has used the same EHR since. After piloting scribe use in Obstetrics & Gynecology for several years, OHSU implemented a university-wide scribe program in 2015. Rather than employ scribes through an external scribe service, the program is managed by OHSU and primarily employs students seeking additional medical experience before embarking on further training (e.g., medical school) rather than individuals with prior clinical training (e.g., technicians, medical assistants). This research was approved by OHSU’s Institutional Review Board which granted a waiver of informed consent for analysis of EHR data and metadata.

Data Collection

To examine how notes written with scribe assistance differ from those written without, we first identified which providers worked with scribes during the years of 2017 and 2018. To do so, we obtained a list of all EHR user IDs for scribes employed by OHSU’s scribe program from its inception in 2015 until July 2020. We then used this list to identify all outpatient encounters in 2017 and 2018 where a scribe edited the patient’s record. From this list of encounters, we identified providers who had worked with a scribe for at least 100 encounters during the study period. To get a sample of notes written by these providers, we leveraged an existing corpus of 200,000 randomly-selected patient records covering the decade from 2009 to 2018. From this corpus we collected all notes written in 2017 and 2018 by providers who regularly worked with scribes as identified above. In addition, we collected demographic data about each encounter such as the billed level of service, patient age, the patient’s assigned sex, and the patient’s self-described race and ethnicity. To look at how notes were written in more detail, we also collected EHR metadata describing how much of each note had been entered via manual typing, copy-pasting, or a note template, as well as which EHR user entered that text. Providers with fewer than 20 scribed notes or 20 non-scribed notes in the corpus were excluded from analysis to ensure a sufficient number of notes of each type for comparison.

Measures

The primary measures employed in this study were note length, the source of note text, and the presence of certain lists of templated patient data. Note length was measured in characters as tracked by the EHR. The source of note text was measured using EHR metadata to count how many of a note’s characters were typed manually, copy-pasted, or inserted with a template. Since this metadata tracks which user entered text via each method, we calculated what proportion of each note’s characters were entered by the primary provider for the encounter, the scribe, or another member of the care team. Finally, we used regular expression matching to identify whether each note contained highly-structured data lists inserted into the note using a template. For example, whether a note contained an imported problem list could be determined by checking if the note contained the text ”Patient Active Problem List” which appears at the start of every list imported using our institution’s standard problem list data link.

Data Analysis

Differences in patient and encounter demographics between scribed and non-scribed visits were evaluated using $\chi^2$ tests. We report mean note lengths and the mean number of characters attributable to each source of note text to ensure component values (i.e., characters of manually typed, templated, and copied text) sum to the average note length. The observed data are not independent (e.g., the lengths of notes written by the same provider are correlated) so we use mixed-effect linear models to assess the association of working with a scribe with 1) note lengths, 2) the number of characters attributable to each text source, and 3) the proportion of notes containing each type of templated patient data list. These models control for the provider as a random effect and the encounter year, patient age, patient assigned sex, patient self-described race, patient self-described ethnicity, and billed level of service as fixed effects. We only report model estimates for the modeled associations with scribe presence as all other model factors were included not to test a specific hypothesis, but to control for potential disparities. Working with a scribe was considered to be significantly associated with an increase or decrease in each of the modeled measures if the model’s p-value for scribe presence was <.05. Variation in how scribes affected different provider’s documentation workflows was also explored by plotting how much longer or shorter scribed notes were than non-scribed notes for each provider. All analyses were performed in Python (v. 3.7.6), using the statsmodels library for statistical modeling.
Results

We identified 89,084 encounters in 2017-2018 where a scribe employed by OHSU’s scribe program had edited the patient’s record. Limiting this list of encounters to office visits, providers who worked with scribes for at least 100 visits over the study period, and visits where only one scribe edited the record left 73,061 potential scribed visits for inclusion in the study. Of these, note data from the randomly sampled records were available for 28,473 scribed visits (39% of potential scribed visits) as well as 24,653 visits conducted by the same providers without a scribe present. This sample excluded 2,718 notes from 14 providers with fewer than 20 scribed or 20 non-scribed notes in the corpus. In total 53,126 visits conducted by 70 providers in 27 different specialties were included in the study (Table 1).

<table>
<thead>
<tr>
<th>Specialty</th>
<th># Providers</th>
<th># Patients</th>
<th># Visits</th>
<th># Scribed Visits</th>
<th># Non-Scribed Visits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obstetrics &amp; Gynecology</td>
<td>16</td>
<td>4,778</td>
<td>13,173</td>
<td>9,008</td>
<td>4,165</td>
</tr>
<tr>
<td>Internal Medicine</td>
<td>6</td>
<td>3,664</td>
<td>8,986</td>
<td>7,446</td>
<td>1,540</td>
</tr>
<tr>
<td>Orthopedics</td>
<td>3</td>
<td>2,134</td>
<td>4,487</td>
<td>2,693</td>
<td>1,794</td>
</tr>
<tr>
<td>Pain Management</td>
<td>6</td>
<td>1,049</td>
<td>4,098</td>
<td>1,497</td>
<td>2,601</td>
</tr>
<tr>
<td>Ophthalmology</td>
<td>2</td>
<td>772</td>
<td>3,451</td>
<td>1,399</td>
<td>2,052</td>
</tr>
<tr>
<td>Otolaryngology</td>
<td>5</td>
<td>1,825</td>
<td>3,239</td>
<td>1,337</td>
<td>1,902</td>
</tr>
<tr>
<td>Dermatology</td>
<td>3</td>
<td>1,773</td>
<td>2,587</td>
<td>429</td>
<td>2,158</td>
</tr>
<tr>
<td>Family Practice</td>
<td>4</td>
<td>1,171</td>
<td>1,880</td>
<td>1,053</td>
<td>827</td>
</tr>
<tr>
<td>Plastic Surgery</td>
<td>3</td>
<td>733</td>
<td>1,429</td>
<td>114</td>
<td>1,315</td>
</tr>
<tr>
<td>18 Other Specialties</td>
<td>27</td>
<td>4,517</td>
<td>9,796</td>
<td>3,497</td>
<td>6,299</td>
</tr>
<tr>
<td>All</td>
<td>70</td>
<td>19,313</td>
<td>53,126</td>
<td>28,473</td>
<td>24,653</td>
</tr>
</tbody>
</table>

Scribed vs Non-Scribed Visits

There were significant differences in patient and encounter demographics for visits documented with and without scribe assistance for which notes were available. Visits with a scribe present were more likely to have been conducted in 2018 rather than 2017 (p<.001), potentially reflecting the increasing use of the scribe program over time, and were more likely to have been billed at a higher level of service (p<.001) (Table 2). Scribed visits were also more likely to have been for older, non-hispanic, and female patients, which may reflect that physicians in Obstetrics and Gynecology were among the first to work with scribes at OHSU and still among the most frequent users of scribes (Table 1).

Note Length

Notes written with the assistance of scribes were significantly longer than those written without scribe assistance. In our sample, the mean note written with a scribe was 6,062 characters long whereas the mean note written without a scribe was 4,430 characters in length. However, this observed difference also reflects differences in the kinds of visits conducted with and without scribes, with some providers and specialties employing scribes at most visits, and others less frequently (Table 1). Controlling for variation in note lengths by provider, patient age, sex, race, ethnicity, encounter year, and billed level of service, notes written with the assistance of scribes were found to be 989 characters longer than those written without scribes (p<.001, 938 to 1,040 95% CI). This represents a 22% increase in length from the average non-scribed note. The association of scribe use with increasing note length varied by provider. While 61 of the 70 observed providers wrote longer notes when working with a scribe—in one case over 4,000 characters longer on average—9 providers wrote shorter notes when working with scribes (Figure 1).

Text Source

The source of note text was significantly different between scribed and non-scribed visits (Figure 2). Controlling for variation in note length by provider, patient age, sex, race, ethnicity, encounter year, and billed level of service, providers wrote an average of 2,024 fewer characters when working with a scribe (p<.001, 1,983 to 2,065 95% CI), and members of the care team other than the provider or the scribe (e.g., residents, technicians, medical assistants) wrote 1,186 fewer characters (p<.001, 1,148 to 1,224 95% CI). Meanwhile, the overall mix of text entered by any
author included significantly more text from all three sources, including 272 more characters of manually typed text ($p < .001, 256 to 289 95\% \text{ CI}$), 112 more of copied text ($p < .001, 86 to 137 95\% \text{ CI}$), and 606 additional characters of templated text ($p < .001, 567 to 644 95\% \text{ CI}$).

Table 2: Select characteristics of scribed and non-scribed visits. P-values from $\chi^2$ tests.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Scribed Visits</th>
<th>Non-Scribed Visits</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient Age (years, median [IQR])</td>
<td>47 [34,65]</td>
<td>44 [29,64]</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Patient, Female (%)</td>
<td>76.0</td>
<td>68.0</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Patient Ethnicity</td>
<td></td>
<td></td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Non-Hispanic</td>
<td>92.0</td>
<td>89.6</td>
<td></td>
</tr>
<tr>
<td>Hispanic</td>
<td>5.2</td>
<td>7.1</td>
<td></td>
</tr>
<tr>
<td>Unknown</td>
<td>2.8</td>
<td>3.4</td>
<td></td>
</tr>
<tr>
<td>Patient Race</td>
<td></td>
<td></td>
<td>&lt; .001</td>
</tr>
<tr>
<td>White</td>
<td>85.2</td>
<td>85.4</td>
<td></td>
</tr>
<tr>
<td>Asian</td>
<td>6.3</td>
<td>4.8</td>
<td></td>
</tr>
<tr>
<td>Black</td>
<td>2.3</td>
<td>2.7</td>
<td></td>
</tr>
<tr>
<td>American Indian/Alaska Native</td>
<td>0.4</td>
<td>0.6</td>
<td></td>
</tr>
<tr>
<td>Native Hawaiians and Other Pacific Islanders</td>
<td>0.3</td>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td>Unknown</td>
<td>5.2</td>
<td>5.6</td>
<td></td>
</tr>
<tr>
<td>Encounter Year (%)</td>
<td></td>
<td></td>
<td>&lt; .001</td>
</tr>
<tr>
<td>2017</td>
<td>43.0</td>
<td>56.5</td>
<td></td>
</tr>
<tr>
<td>2018</td>
<td>57.0</td>
<td>43.4</td>
<td></td>
</tr>
<tr>
<td>Billed Level of Service (%)</td>
<td></td>
<td></td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Level 1</td>
<td>0.1</td>
<td>0.4</td>
<td></td>
</tr>
<tr>
<td>Level 2</td>
<td>3.5</td>
<td>11.0</td>
<td></td>
</tr>
<tr>
<td>Level 3</td>
<td>25.0</td>
<td>25.5</td>
<td></td>
</tr>
<tr>
<td>Level 4</td>
<td>36.4</td>
<td>25.1</td>
<td></td>
</tr>
<tr>
<td>Level 5</td>
<td>7.1</td>
<td>6.3</td>
<td></td>
</tr>
<tr>
<td>Other$^a$</td>
<td>28.0</td>
<td>31.6</td>
<td></td>
</tr>
</tbody>
</table>

$^a$ Includes encounters without level of service specified, or billed with a code lacking the typical 1-5 office visit code levels.

Figure 1: Difference in mean note length between notes written with and without scribes, by provider.
Figure 2: Mean number of note characters by source of note text and author for notes written with and without scribes.

**Templated Data Lists**

Scribed notes were observed to include certain highly structured lists of patient data more often. For example, scribed notes were more than twice as likely to include an imported patient problem list (29.5 vs. 12.6% of notes) and 10 times as likely to include a list of reviewed lab results (11.1 vs. 1.1%). However, these differences as well as the difference in how often notes included a structured allergy list (47.8 vs. 36.5%) were not significant when controlling for factors such as the provider, patient age, sex, race, ethnicity, encounter year, and billed level of service. That is, the observed differences in how often these items were included in the note were driven more by the changing practices of a few providers than being a consistent difference across all providers. By contrast, scribed notes were significantly more likely to include medication lists as well as structured lists of past medical, past surgical, and family history (Table 3).

Table 3: Percent of scribed and non-scribed notes containing certain templated lists of patient data. Parameter estimates and p-values are given for the scribe factor in mixed-effect models of each list’s presence.

<table>
<thead>
<tr>
<th>List of Data</th>
<th>% of Scribed Notes</th>
<th>% of Non-Scribed Notes</th>
<th>Modelled Difference (%) (Est. [95% CI])</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Problem List</td>
<td>29.5</td>
<td>12.6</td>
<td>0.0 [-0.5 to 0.4]</td>
<td>.893</td>
</tr>
<tr>
<td>Medication List</td>
<td>47.5</td>
<td>41.5</td>
<td>6.5 [5.8 to 7.2]</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Allergy List</td>
<td>47.8</td>
<td>36.5</td>
<td>-0.1 [-0.7 to 0.6]</td>
<td>.834</td>
</tr>
<tr>
<td>Past Medical History</td>
<td>45.0</td>
<td>37.8</td>
<td>4.6 [3.8 to 5.4]</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Past Surgical History</td>
<td>40.5</td>
<td>32.8</td>
<td>6.6 [5.8 to 7.4]</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Family History</td>
<td>34.9</td>
<td>20.4</td>
<td>6.9 [6.2 to 7.6]</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Labs Reviewed</td>
<td>11.1</td>
<td>1.1</td>
<td>0.1 [-0.2 to 0.4]</td>
<td>.452</td>
</tr>
</tbody>
</table>

Variation by Provider

Lastly, to examine variation by provider and specialty, we plotted the mean difference in note length, both overall and by text source, for the nine providers with the greatest increase and the greatest decrease in note lengths when working with a scribe, as well as all providers in the four specialties with the most scribed encounters (Figures 3 & 4). Providers did not display a consistent trend but varied in whether most of the additional (or decreased) text was manually typed, templated or copied. For example, while several providers had large increases in templated and manually typed text, for other providers the increase in note length was almost entirely attributable to additional copied text. While trends were more apparent when providers were grouped by specialty, not all providers in the same specialty exhibited the same trends (Figure 4).
**Figure 3:** Average (mean) difference in source of note text for providers with the largest increase and largest decrease in note length between notes written with and without scribe assistance.

**Figure 4:** Average (mean) difference in source of note text between notes written with and without scribe assistance for providers in the four specialties with the most study encounters.
Discussion

This study was designed to test two primary hypotheses: 1) that scribed notes were longer than non-scribed notes and 2) that scribed notes contained more manually entered text, but not more templated or copied text. In addition, it began to explore the variation in how providers adapted their documentation workflows to working with a scribe.

Longer Scribed Notes

We find strong support for the first hypothesis. After controlling for variation by patient, provider, and encounter characteristics, we find scribed notes were 989 characters longer than those written without scribe assistance—equivalent to about 150 words of additional text. There are a number of reasons why scribed notes may be longer, including scribes manually typing more detailed patient histories or including templated attestations of their contribution to the note. Scribed notes may also be capturing more detailed information to support higher levels of billing, though the models employed in this study already control for the propensity of scribed visits to be billed at higher levels of service. Whatever the precise cause, the observed difference in note length raises the concern that while scribes may save physician time, they may also contribute to note bloat. While 1,000 extra characters in a single note may seem a small increase, it translates to a 22% increase in length over the average non-scribed note. Multiplied across the 20 or more notes a provider may write in a single day—or that another physician may need to scan during chart review—the accumulated additional text may profoundly affect EHR burden. With clinical notes in the United States already 4-times longer than those in other countries, care must be taken not to bloat notes further.

Source of Text in Scribed Notes

Beyond observing differences in note lengths, this study was also designed to investigate how text was entered into scribed and non-scribed notes. We find evidence supporting part of our second hypothesis; that scribed notes would contain more manually entered text, but not more templated or copied text. Scribed notes did indeed contain more manually typed text (272 characters), but also more copied and templated text (112 and 606 additional characters respectively). The overall difference in lengths between scribed and non-scribed notes thus appears to be driven more by note templates than manually typed text. And despite the risks of copying note text to patient safety and note quality, we see an increase in the amount of copied note text when working with a scribe. Invoking templates and copy-pasting prior notes may help scribes document consistently, and quickly adapt to the preferences of different providers. However it may also propagate outdated or erroneous findings, as when copying a prior assessment and plan, or using a template to generate a review of systems with default findings that were not actually observed.

Trends were also apparent in shifting note authorship. As might be expected, providers wrote substantially less text in scribed notes, entering fewer than half the characters they would in a typical note written without scribe assistance. Yet, they still entered a non-trivial amount of text. Scribes also contributed substantial amounts of text, with just the scribe contribution to the average scribed note being longer than the entire average non-scribed note. We also observed substantially less text being entered into scribed notes by authors other than the provider or scribe. While data are not presented here, a post-hoc analysis revealed at least part of this drop was due to medical residents writing large portions of non-scribed notes for some providers, but rarely doing so when a scribe was present. This may reflect some providers not electing to employ scribes during resident-led clinics. Beyond demonstrating changing patterns of note authorship, these data highlight the highly collaborative nature of note writing during both scribed and non-scribed visits.

Variation Amongst Providers and Specialties

In addition to testing our two primary hypotheses, this study was also designed to observe variation in how providers adapted their documentation workflows to working with a scribe. Figures 3 & 4 demonstrate the varying impact, even within specialties, that working with a scribe had on how providers drafted their notes. While some providers included substantially more templated and manually written text when working with a scribe, others included dramatically more copied text but only marginally more (or sometimes less) manually typed or templated text. Still other providers reduced the amount of text included in their notes from all three sources of manual typing, copying, and templates. There are likely many individual changes to documentation workflows underlying these observed differences. Providers with large increases in copied note text for scribed visits may have adopted a workflow where the
scribe copies entire sections of prior notes (e.g., history of present illness, assessment and plan) and pastes them into
the current note. Dramatic increases in the amount of templated text may be a result of providers creating entirely new
note templates for scribes to use which import more data lists such as medications or past medical history to help the
provider review the patient’s case.

Future work is needed to observe differences in documentation workflows more directly and to determine which ways
of working with scribes are most efficient, effective, and safe.23 There is existing evidence that provider’s individual
preferences, differences in EHR design, and having multiple ways to enter the same information can drive variation in
provider’s documentation.24 These differences can lead to variation in the completeness of patient records, and may
contribute to patient harm if data are incorrect or difficult to find. While more research on scribed documentation
workflows is needed, it is clear from the evidence collected in this study that working with a scribe does not have one
consistent impact on documentation workflows, but varies from provider to provider.

Implications for Research, Training, and Design

There is existing evidence that provider’s individual
preferences, differences in EHR design, and having multiple ways to enter the same information can drive variation in
provider’s documentation.24 These differences can lead to variation in the completeness of patient records, and may
contribute to patient harm if data are incorrect or difficult to find. While more research on scribed documentation
workflows is needed, it is clear from the evidence collected in this study that working with a scribe does not have one
consistent impact on documentation workflows, but varies from provider to provider.

Future work is needed to observe differences in documentation workflows more directly and to determine which ways
of working with scribes are most efficient, effective, and safe.23 There is existing evidence that provider’s individual
preferences, differences in EHR design, and having multiple ways to enter the same information can drive variation in
provider’s documentation.24 These differences can lead to variation in the completeness of patient records, and may
contribute to patient harm if data are incorrect or difficult to find. While more research on scribed documentation
workflows is needed, it is clear from the evidence collected in this study that working with a scribe does not have one
consistent impact on documentation workflows, but varies from provider to provider.

Implications for Research, Training, and Design

There is existing evidence that provider’s individual
preferences, differences in EHR design, and having multiple ways to enter the same information can drive variation in
provider’s documentation.24 These differences can lead to variation in the completeness of patient records, and may
contribute to patient harm if data are incorrect or difficult to find. While more research on scribed documentation
workflows is needed, it is clear from the evidence collected in this study that working with a scribe does not have one
consistent impact on documentation workflows, but varies from provider to provider.

Implications for Research, Training, and Design

There is existing evidence that provider’s individual
preferences, differences in EHR design, and having multiple ways to enter the same information can drive variation in
provider’s documentation.24 These differences can lead to variation in the completeness of patient records, and may
contribute to patient harm if data are incorrect or difficult to find. While more research on scribed documentation
workflows is needed, it is clear from the evidence collected in this study that working with a scribe does not have one
consistent impact on documentation workflows, but varies from provider to provider.

Implications for Research, Training, and Design

There is existing evidence that provider’s individual
preferences, differences in EHR design, and having multiple ways to enter the same information can drive variation in
provider’s documentation.24 These differences can lead to variation in the completeness of patient records, and may
contribute to patient harm if data are incorrect or difficult to find. While more research on scribed documentation
workflows is needed, it is clear from the evidence collected in this study that working with a scribe does not have one
consistent impact on documentation workflows, but varies from provider to provider.

Implications for Research, Training, and Design

There is existing evidence that provider’s individual
preferences, differences in EHR design, and having multiple ways to enter the same information can drive variation in
provider’s documentation.24 These differences can lead to variation in the completeness of patient records, and may
contribute to patient harm if data are incorrect or difficult to find. While more research on scribed documentation
workflows is needed, it is clear from the evidence collected in this study that working with a scribe does not have one
consistent impact on documentation workflows, but varies from provider to provider.

Implications for Research, Training, and Design

There is existing evidence that provider’s individual
preferences, differences in EHR design, and having multiple ways to enter the same information can drive variation in
provider’s documentation.24 These differences can lead to variation in the completeness of patient records, and may
contribute to patient harm if data are incorrect or difficult to find. While more research on scribed documentation
workflows is needed, it is clear from the evidence collected in this study that working with a scribe does not have one
consistent impact on documentation workflows, but varies from provider to provider.

Implications for Research, Training, and Design

There is existing evidence that provider’s individual
preferences, differences in EHR design, and having multiple ways to enter the same information can drive variation in
provider’s documentation.24 These differences can lead to variation in the completeness of patient records, and may
contribute to patient harm if data are incorrect or difficult to find. While more research on scribed documentation
workflows is needed, it is clear from the evidence collected in this study that working with a scribe does not have one
consistent impact on documentation workflows, but varies from provider to provider.

Implications for Research, Training, and Design

There is existing evidence that provider’s individual
preferences, differences in EHR design, and having multiple ways to enter the same information can drive variation in
provider’s documentation.24 These differences can lead to variation in the completeness of patient records, and may
contribute to patient harm if data are incorrect or difficult to find. While more research on scribed documentation
workflows is needed, it is clear from the evidence collected in this study that working with a scribe does not have one
consistent impact on documentation workflows, but varies from provider to provider.

Implications for Research, Training, and Design

There is existing evidence that provider’s individual
preferences, differences in EHR design, and having multiple ways to enter the same information can drive variation in
provider’s documentation.24 These differences can lead to variation in the completeness of patient records, and may
contribute to patient harm if data are incorrect or difficult to find. While more research on scribed documentation
workflows is needed, it is clear from the evidence collected in this study that working with a scribe does not have one
consistent impact on documentation workflows, but varies from provider to provider.

Implications for Research, Training, and Design

There is existing evidence that provider’s individual
preferences, differences in EHR design, and having multiple ways to enter the same information can drive variation in
provider’s documentation.24 These differences can lead to variation in the completeness of patient records, and may
contribute to patient harm if data are incorrect or difficult to find. While more research on scribed documentation
workflows is needed, it is clear from the evidence collected in this study that working with a scribe does not have one
consistent impact on documentation workflows, but varies from provider to provider.
Conclusion

This study demonstrates a strong association between working with scribes and writing longer notes. After controlling for factors such as patient, provider, and encounter characteristics, scribed notes were found to have 989 more characters than notes written without scribe assistance, or roughly 150 additional words. While a portion of this additional text came from manually typed text, the majority came from additional templated text. This study also demonstrates substantial variation in how providers adapted their documentation workflows when working with scribes, as demonstrated by changes to the amount of text incorporated into their notes from manual typing, templates, and copy-pasting. These results suggest working with scribes may contribute to note bloat, but that provider’s individual documentation workflows, including the design of their note templates, may have a large impact on the contents of scribed notes and ultimately the accuracy, quality, and utility of scribed documentation.

Acknowledgements

This research was supported by grants R00LM12238, T15LM007088, and P30EY10572 from the National Institutes of Health (Bethesda, MD), grant R01HS025141 from the Association of Healthcare Research and Quality (Rockville, MD), and by unrestricted departmental funding from Research to Prevent Blindness (New York, NY).

References

Design of digital walking programs that engage prostate cancer survivors: Needs and preferences from focus groups

Savitha Sangameswaran, B.Tech1, Courtney Segal, BA1, Dori E Rosenberg, PhD MPH2, Reggie Casanova-Perez, MS1, David Cronkite, MS2, John L. Gore, MD MS, 1 Andrea L. Hartzler, PhD1

1University of Washington, Seattle, WA 2 Kaiser Permanente Washington Health Research Institute Seattle, WA.

Abstract

The majority of prostate cancer survivors do not meet physical activity (PA) recommendations. Although technology has shown to promote PA, engagement has been a challenge. This mixed method study characterizes survivors’ needs and preferences for digital walking programs. Through focus groups and surveys, we engaged prostate cancer support groups to describe PA motivators and barriers, interest in improving PA, and preferences for design features of a future digital walking program. Identified motivators (peers, positive thinking) and barriers (health issues) reflect PA needs that impact engagement. The most preferred features include: (1) well-curated, specific content, (2) individualized feedback from trusted sources, (3) moderated peer discussion, and (4) support from small teams and peer mentors. These findings inform digital PA programs that survivors will find engaging and can promote PA.

Introduction

Prostate cancer is the second most common cancer in men with 1 in 8 diagnosed during his lifetime.1 With an estimated 248,530 new cases of prostate cancer in 20211 and a 5-year relative survival rate of 98%2, more than 3.1 M men in the United States are prostate cancer survivors3. Given the high disease burden of prostate cancer there is a need to identify strategies to improve quality of life among cancer survivors. Regular physical activity (PA) improves fitness, vitality and quality of life among cancer survivors4-6, delays cancer progression7, and reduces the risk of prostate cancer mortality8. Despite the protective benefits of PA for prostate cancer survivorship 85% of prostate cancer survivors report not meeting PA recommendations9.

Even though traditional supervised exercise programs show effectiveness in improving PA for prostate cancer survivors,10 barriers including cost, lack of time and support, access to facilities, and exercise partners, as well as disease-specific barriers like treatment side effects (e.g., urinary incontinence, fatigue) can limit engagement in these programs11-12. There is an opportunity to leverage engaging digital technologies in distance-based programs that bring support to survivors in their own neighborhoods and communities as a feasible low cost alternative.

Digital PA programs provide the opportunity to deliver behavior change techniques13,14 (e.g., goal setting, feedback, social support) via an easy to access, low cost and scalable modality. Digital PA programs that use technologies like fitness trackers, mobile phones, social media have been shown to promote PA among prostate cancer survivors15-17. Yet, lack of adherence and sustained engagement in these digital interventions by prostate cancer survivors limits the effectiveness of these interventions17-19. Thus, a major gap in prior research is how to design digital PA programs that prostate cancer patients will find engaging and want to use. There is a need to understand what type of digital technology appeals most to prostate cancer survivors and how to package existing behavior change features in these digital tools to promote PA.

To address this gap, we engaged prostate cancer survivors and their loved ones in focus groups to understand their needs and preferences for digital walking interventions. We particularly focused on walking as it is one of the most accessible forms of PA that requires no training and might be sustainable throughout a patient’s life. The purpose of this study was to characterize the user needs and design preferences of digital walking programs that leverage fitness trackers and social media features for prostate cancer survivors. Findings offer guidance to inform the design of supportive PA content, self-tracking features, and peer support elements of digital tools that can promote walking.

Methods

We conducted a concurrent mixed methods20 study with focus groups and surveys. Study procedures were approved by the University of Washington Institutional Review Board. We recruited from prostate cancer support groups in the Seattle metropolitan area from July to August 2018 for focus groups conducted at regularly scheduled support group meetings. Each focus group lasted for 60-90 minutes, was moderated by a team member and comprised 3 parts.
First, the moderated led group discussion on perceived benefits and barriers of walking. Participants wrote down ideas to the question “If you had a magic wand what is that one thing that would help you walk more?”. The moderator collected these “magic wand” ideas marked on index cards and read examples aloud to ground group discussion on PA motivators and barriers. This discussion helped to establish and characterize user needs for digital PA programs.

Second, the facilitator introduced the design concept of a digital walking program that incorporated a fitness tracker (i.e., Fitbit) and a social media-based platform (i.e., Facebook) to increase step count. The features of the digital walking program were designed based on best practice and evidence-based recommendations\textsuperscript{13,14,21} to facilitate personalized goal tracking, feedback, and social support. We collected qualitative data on design features participants prefer across 4 sets of options presented through mock interfaces: content, PA feedback, group discussion, and teamwork (Figure 1). Participants discussed likes and dislikes about these feature and ways each could be improved.

Third, after the focus group, participants completed an optional exit survey to report participant characteristics (i.e., demographics, technology experience), interest in improving PA (not at all interested, somewhat interested, very interested), and importance of each design feature (not important, less important, more important, most important).

![Figure 1. Mock interfaces for design features: A) Content, B) PA feedback, C) group discussion, and d) teamwork](image)

**Data analysis**

Focus group sessions were audio recorded and transcribed for qualitative analysis. Two members of the research team took notes during each session to supplement transcripts. We deductively coded this data with a high-level a priori coding schema: PA motivators, PA barriers, and preferences for each of the 4 features (i.e., content, PA feedback, group discussion, and teamwork). A team member (SS) analyzed transcripts using the coding schema in Dedoose\textsuperscript{22}. The researchers (AL and SS) met during the coding process to discuss sub-codes and refine coding categories.

We quantitatively analyzed survey data with SPSS 26.0 (SPSS Inc). We used descriptive statistics to summarize participant characteristics, interest in using digital PA programs, and preferred design features based on ratings of importance. We used inferential statistics to compare interest level and preferences for design features across all participants and among the 3 focus group sessions. We used Friedman tests ($X^2$) to assess differences across participants and Kruskal Wallis tests (H) to assess differences among groups. We used Dunn tests (Z)\textsuperscript{23} with a Bonferroni adjustment for post hoc pairwise comparisons for significant results.

**Results**

1. Participants
Across the 3 focus groups (FG1, FG2, FG3), we engaged 61 members of prostate cancer support groups, of whom 49 responded to the survey (80%). Table 1 summarizes the characteristics of survey respondents who were predominantly older white males, college educated or greater, and 44 reported having been diagnosed within 4 months to 21 years (mean = 7 SD=6). Respondents who were prostate cancer survivors indicated a range of treatments (e.g., radiation, cryotherapy, cyber knife). Of the 5 respondents who did not report having a prostate cancer diagnosis, three identified as a spouse, and one as a caregiver, and one did not report their role. The majority of respondents reported using the internet and smartphones, but less than half reported using social media and less than one quarter reported using a fitness tracker.

### Table 1. Characteristics of survey respondents

<table>
<thead>
<tr>
<th></th>
<th>All (n=49)</th>
<th>FG 1 (n=20)</th>
<th>FG 2 (n=11)</th>
<th>FG 3 (n=18)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mean (sd)</strong></td>
<td>Range</td>
<td>N (%)</td>
<td>Mean (sd)</td>
<td>N (%)</td>
</tr>
<tr>
<td><strong>Age</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>72(7)</td>
<td>59-89</td>
<td>72(5)</td>
<td>60-79</td>
</tr>
<tr>
<td><strong>Sex</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>45 (92)</td>
<td>18 (90)</td>
<td>11 (100)</td>
<td>16 (89)</td>
</tr>
<tr>
<td>Female</td>
<td>4 (8)</td>
<td>2 (10)</td>
<td>0</td>
<td>2 (11)</td>
</tr>
<tr>
<td><strong>Race</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>43 (88)</td>
<td>18 (90)</td>
<td>9 (82)</td>
<td>16 (89)</td>
</tr>
<tr>
<td>Non-white</td>
<td>4 (8)</td>
<td>1 (5)</td>
<td>2 (18)</td>
<td>1 (6)</td>
</tr>
<tr>
<td>Declined to state</td>
<td>2 (4)</td>
<td>1 (5)</td>
<td>0</td>
<td>1 (6)</td>
</tr>
<tr>
<td><strong>Ethnicity</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hispanic/Latino</td>
<td>0 (0)</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Not Hispanic/Latino</td>
<td>43 (88)</td>
<td>17 (85)</td>
<td>9 (82)</td>
<td>17 (94)</td>
</tr>
<tr>
<td>Declined to state</td>
<td>6 (12)</td>
<td>3 (15)</td>
<td>2 (18)</td>
<td>1 (6)</td>
</tr>
<tr>
<td><strong>Education</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High school</td>
<td>9 (18)</td>
<td>3 (15)</td>
<td>2 (18)</td>
<td>4 (22)</td>
</tr>
<tr>
<td>College</td>
<td>17 (35)</td>
<td>6 (30)</td>
<td>4 (36)</td>
<td>7 (39)</td>
</tr>
<tr>
<td>Advanced degree</td>
<td>20 (41)</td>
<td>9 (45)</td>
<td>5 (45)</td>
<td>6 (33)</td>
</tr>
<tr>
<td>Doctorate</td>
<td>2 (4)</td>
<td>2 (10)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Decline to state</td>
<td>1 (2)</td>
<td>0</td>
<td>0</td>
<td>1 (6)</td>
</tr>
<tr>
<td><strong>Technology use</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Use Internet</td>
<td>46 (94)</td>
<td>20 (100)</td>
<td>10 (91)</td>
<td>16 (89)</td>
</tr>
<tr>
<td>Use smartphone</td>
<td>37 (76)</td>
<td>14 (70)</td>
<td>9 (82)</td>
<td>14 (78)</td>
</tr>
<tr>
<td>Use social media</td>
<td>19 (39)</td>
<td>8 (40)</td>
<td>2 (18)</td>
<td>9 (50)</td>
</tr>
<tr>
<td>Use a fitness tracker</td>
<td>11 (22)</td>
<td>7 (35)</td>
<td>0</td>
<td>4 (22)</td>
</tr>
</tbody>
</table>

2. Motivators and Barriers to PA

Across focus groups a range of “magic wand” ideas surfaced. Table 2 summarizes a few of these ideas. We categorized ideas as technology-based (e.g., reminders, virtual competitions), peer support (e.g., walking buddy, coach), and other ideas, such as “a zapper”.

As the group discussed these magic wand ideas, participants described individual experiences with PA and key PA motivators and barriers surfaced, which we summarize below.

**PA motivators:** Participants expressed both internal and external motivators for PA. Participants told us that considering PA as likable and fun is an important internal motivator for PA. For example:

“Let’s start with basics – what do you like to do? I like to go for a walks and runs in forests, I like to move in beautiful places….If you are happy when you are walking, you are in magic land. So that’s what I do. I walk the trails and the cliffs… and I wouldn’t miss it for the world. So that’s like we get things turned around just get the tools that makes it work. What works for us and what’s suits our nature.” (FG 3)

“Nobody likes boring exercises! It needs to be made fun and it could be anything from dance classes to you know going on group walks or things like that….rather than that just going to the gym and going for a walk there are...
...a lot of other things you can do to get the exercise that are a lot more fun at the same time. Sometimes people don’t connect fun activities with exercise you don’t even know. But it is.” (FG 3)

Table 2: “Magic Wand” ideas from participants

<table>
<thead>
<tr>
<th>Technology-based ideas</th>
<th>Peer support ideas</th>
<th>Other ideas</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Use multiple alarm each day</td>
<td>• Find a walking buddy</td>
<td>• Good weather</td>
</tr>
<tr>
<td>• Virtual community to provide social support</td>
<td>• Form interest groups that meet and walk at the same time</td>
<td>• Problem – Motivation</td>
</tr>
<tr>
<td>• Reward (s) for meeting goals.</td>
<td>• Friend (or Tech) ready to Pace Me</td>
<td>• Hypnotherapy</td>
</tr>
<tr>
<td>• Notification if you are not moving more than a “vibrate”</td>
<td>• A coach (with more information on exercise plans)</td>
<td>• A zapper if inactive for more than 24 hours</td>
</tr>
<tr>
<td>• Virtual competition</td>
<td>• An accountability partner</td>
<td>• Watch sports TV for only as many minutes as exercised that day</td>
</tr>
<tr>
<td>• Reminder system on phone</td>
<td>• Blue zones in our community</td>
<td>• After achieving goal the magic door opens to a nice bowl of ice cream</td>
</tr>
<tr>
<td>• Weight scale</td>
<td>• Meet at coffee shop and support online via social involvement</td>
<td>• Do something you enjoy doing</td>
</tr>
<tr>
<td>• Feedback as to length, intensity of exercise, app phone etc.</td>
<td></td>
<td>• Club membership</td>
</tr>
<tr>
<td>• Show people available technology today</td>
<td>• Software and support devices to do self-evaluation of overall fitness</td>
<td>• Brain implant to motivate</td>
</tr>
<tr>
<td>• Software and support devices to do self-evaluation of overall fitness</td>
<td>• Audio reminder for exercise</td>
<td>• USB port on body</td>
</tr>
<tr>
<td>• Audio reminder for exercise</td>
<td>• Find a walking buddy</td>
<td>• Make a machine to kick my back side</td>
</tr>
</tbody>
</table>

Participants also indicated that personal characteristics, like fear of failure, serve as internal motivators:

“...I hated failing. That [failure] was more of a push than the rewards for being successful.” (FG2)

Another participant expressed the value of positive attitude and thinking as an internal PA motivator:

“...Positive attitude is probably the basic for my good health... There are a lot of opportunities for good health like health clubs, there are things that people can do that don’t get enough exposure. People get on a positive line of thought and those problems start to get away if you start doing what’s best for you.” (FG 3)

In addition to internal motivators, participants also indicated external motivators from others. Advice from healthcare providers was a key motivator for many, for example, participants shared:

“My doctor told me that I can start hormone therapy and one of the side effects is that it weakens your structural system, and so you need to do some basic strength exercise to combat that down. Well that was the only motivation I needed, it’s just that.” (FG 2)

“I asked the surgeon “what’s the best exercise I can do”? He said “Walking. You need to walk”. And then he added “Not just walk but you’ve got to press yourself.” (FG 2)

Peers were also described as external motivators for PA. For example, participants told us:

“The people that I know, or I’ve come to know, at the YMCA that work out at the same time I do. And they’ve become friends. And I look forward to seeing them... they are friends. I feel like letting them down if I didn’t [go].” (FG 2)

“I think the strongest motivation or motivator would be other people in the group. I don’t know if one of them is not feeling well, the others might go “we missed you.” (FG 1)

Finally, participants indicated the importance of family for motivating them to get more PA:

“My wife is on Fitbit she gets these badges for walking across and stuff like that. She’s competed against other family members for a while. There was a kind of competition and things you know.” (FG 2)

PA barriers: Participants focused on health factors that negatively impact PA, such as recovering from medical procedures or health conditions. For example, participants told us:

“When you have something happen like surgery or an acute event it’s like you lose a lot of ground and need a lot of support on the way back.” (FG 3)

1072
“if someone has irritable bowel it makes it a challenge to head outside without having facilities nearby.” (FG 1)

3. Interest in improving PA

Figure 2 summarizes interest level of survey respondents by focus group. Across all participants, there was high interest in “Improving my PA level” (mean=2.6, SD=0.6), but less interest in “Using technology to improve my PA level” (mean=2.0, SD=0.6), “Using a Fitbit tracker to improve PA level” (mean=1.8, SD=0.7), or “Using a Facebook group to improve my PA level” (mean=1.5, SD=0.6). Across all participants, this interest varied significantly ($\chi^2 (3) = 77.50, p<0.001$). Post hoc pairwise comparisons show significantly greater interest in “Improving my PA level” than improving PA level by (1) using technology ($Z=0.967, p<0.002$), (2) using Fitbit ($Z=1.272, p<0.001$), or (3) using a Facebook group ($Z=1.804, p<0.001$). We also found significantly greater interest across all participants in using technology to improve PA level than using a Facebook group to improve PA level ($Z=0.837, p=0.011$).

Figure 2: Interest in four options for improving PA by focus group (1 = “Not at all interested”; 3 = “Very Interested”)

Among the three focus groups, there was a significant difference in interest for using a Facebook group to improve PA level ($H(2) = 11.367, p = 0.003$). Post hoc pairwise comparison indicated this difference was between FG2 and FG3 ($Z=15.167, p = 0.004$). There were no other significant differences among focus groups.

Group discussion supported these differences and provided insight into participants’ somewhat guarded interest in technologies like Fitbit trackers and Facebook. Use of these technologies was generally low (Table 1) but not universal. One participant expressed challenges with using technology for PA because:

“The technology changes so quickly” (FG 3)

Other participants shared negative experiences and challenges with wearable devices:

“I had a Fitbit and it really helped me walk. But it fell apart in less than a year.” (FG 1)

“If you ride a bike for an hour, you get this many calories burned. If you walk for an hour, you get this many calories burned - when it is quite different when you’re going up that big hill or down that big hill. So, I like a little bit more finding ways of getting a little bit more specific.” (FG 2)

Some participants who were not social media users expressed disinterest:

“I don’t want to participate in a Facebook which I don’t belong to now.” (FG 2)

However, other participants liked the idea of using technology to track PA data, schedule PA, and for reminders. Participants also expressed interest in using technology for self-monitoring:
“So you can do self-evaluation of where you are at... A bunch of data there that he [referring to a coach] compares to your age group. If there was something similar where you could do your own self-evaluation will be helpful.” (FG 3)

Some participants also liked being part of an online group. For example, one participant stated:

“Even in this group I mean there is probably 6 or so different categories of pace or style or even interest. I mean if we can network with real people, that’s the motivation that people like this go for I think...” (FG1)

4. Preferred design features of digital walking programs

Figure 3 summarizes survey respondents’ mean importance ratings for the nine design options categorized by content, PA feedback, discussion and teamwork. Across all participants, there was a significant difference in perceived importance across the nine design options ($X^2 = 126.301, p <0.001$). Post hoc comparisons (Table 2) indicate that some features, such as “accessing posts with educational information”, were rated significantly more important than “receiving feedback via posts from group members” ($z=2.024, p=0.029$), “starting my own discussion” ($z=3.024,p<0.001$), “competing for badges” ($z=4.866,p<0.001$), “comparing step count on a leaderboard” ($z=3.866,p<0.001$), and “working with a team towards a step goal” ($z=2.813, p=0.011$). In contrast, “competing for badges” was rated less important than all other options except “starting my own discussions” ($Z=3.044 p=0.084$).

![Figure 3: Participant importance of social media features by focus group](chart)

We describe quantitative differences and qualitative comments that support differences within the four sets of options categorized by content, PA feedback, discussion and teamwork below.

**Table 2:** p values of pairwise comparisons of content options across all survey respondents (n=49)
Across all participants, the highest rated feature on average was “accessing posts with educational content” (mean = 3.2 out of 5, SD= 0.7). However, there were a not significant difference between importance of this option and “receiving supporting posts from peers” (mean = 2.6, SD= 0.9). Among focus groups, there was a significant difference in importance of “receiving supportive posts from peers” (H(2) = 14.615, p < 0.001). Post hoc comparisons found this difference was between FG2 and FG3 (z= -3.141, p = 0.005) and between FG1 and FG3 (z= -3.354, p = 0.002).

Focus group discussion helps explain and further define these design preferences. Participants indicated that they preferred receiving authoritative information from researchers, coaches, mentors, and clinicians. They expressed interest in receiving quality information that was well curated. Participants also expressed greater interest in receiving information that was specific to prostate cancer than general PA information.

“Well curated, authoritative, interesting content is key.” (P48, FG3)

“When I first got diagnosed with prostate cancer that’s what I did it [referring to finding information on google] and you get all kinds of information too much, perhaps if it was something like this that’s more directed it could be more useful” (FG 3)

Other types of information participants preferred related to personalized exercise, activities, healthy diet and recipes, and goal setting.

**PA feedback options:** Overall, participants rated the importance of “receiving feedback via private messages from the researcher” (mean = 2.7, SD=1.0) higher than “receiving feedback via posts from group members” (mean = 2.3, SD =0.9), but this difference was not significant. Among focus groups there was a significant difference in importance of “receiving feedback via posts from group members” (H(2) = 11.076, p = 0.004). Post hoc comparisons found this difference was between FG2 and FG3 (Z=-3.302, p = 0.003).

Through discussion, participants further describe these preferences. Some indicated that rather than receiving PA feedback from strangers (e.g., people on social media they do not know) they would prefer receiving feedback either from an “accountability partner”, such as a “moderator”, “coach”, or “matched peer”. Participants expressed a preference for in-person meetings to get to know other participants, such a peer support person, before joining a social
media-based PA program. Some felt that meeting in-person could help build trust, shared goals, and motivate progress toward PA goals.

“So it should be someone we know, we like and trust - it takes about a meeting or two probably.”(FG1)

“I think the strongest motivation or motivator would be other people in the group. I don’t know if one of them is not feeling well the other might go “we missed you”.” (FG1)

Some participants were also open to receiving automated feedback from technology:

“Just some kind of feedback that ‘You made it!’ doesn’t need to come from a person”(FG 2)

**Discussion options:** Across all participants, the importance of “having a researcher moderate discussions” (mean = 2.7, SD =0.9) was rated significantly higher than “starting my own discussions” (mean = 2.0, SD =0.8) (z=3.286, p=0.037). There were no significant differences among focus groups for these two discussion options.

Despite the quantitative preference for moderated over self-initiated discussions from surveys, focus group discussion indicated that participants preferred a combination of moderated and peer discussion. They were interested in discussing with a group they were familiar with who shared a similar situation. Many participants indicated that the support group they were currently a part of provided them with opportunities for discussion and a similar group discussion online might be similarly beneficial.

“We can always use more data and information and that’s the stuff we would like to see. Research thoughts about what we can do better and mitigate any of the side effects and to work on not having again those kind of stuff but food not to eat, what foods to eat more of, dietary concerns if those things are proven to be effective and helpful.” (FG2)

**Teamwork options:** Across all participants, the importance of “working with a team toward step goals” (mean = 2.3, SD = 1.0) was rated higher than “comparing your step count on a leaderboard” (mean = 1.6, SD=0.8), which was rated higher than “competing for badges” (mean = 1.2 , SD=0.4). There was a significant difference between competing for badges and working with a team towards a step goal (Z = -4.436, p<0.001). There was no significant difference among focus groups for any of the 3 teamwork options.

Focus group discussion further describes these preferences. Most participants were not interested in competing with one another, which they attributed to age. Some participants indicated they were a little competitive, while most others indicated that they were only interested in competing with themselves and not against others. Thus, participants expressed little interest in competing for badges or comparing goals on a leaderboard for rewards and incentives.

“I could care less if John Doe who I don’t know has 10 million steps doesn’t do me.. it's not going to motivate me any further”” (FG 2)

“FGP (female) - I think competing for badges is not for our age group maybe for younger age group but just did you make your goal yeah did you set a goal did you make your goal if you did give yourself a pat on the back if you did great” (FG 3)

In contrast to competition, rewards, and incentives, participants expressed much greater interest in working with a team towards PA goals, sharing experiences, mentoring each other and motivating each other. Participants preferred closed groups with opportunities to socialize, learn from each other, share information, and find exercise partners or walking buddies.

“The only part about the teams, the last one, ‘working with the team towards step goals’ is if there is a way to set up a team where we can help each other and things more kind of like a mentorship type stuff if somebody is really struggling is there a role I can play to help that person in that process by being a team member with them and maybe that would help them. I might look at that.” (FG 2)

**Discussion**

Few studies have examined prostate cancer survivors’ needs and design preferences for digital walking programs, despite their potential. Through this concurrent mixed methods study, members of prostate cancer support groups offered valuable input that informs development of digital walking programs that have potential to engage users and thus promote PA. The qualitative data from the focus groups helped with interpretation of the quantitative survey data and provided context to the results. Findings are important because prostate cancer survivors need PA support and experience barriers to program access and meeting national PA recommendations. Yet there has been little in the way
of design guidance for engaging digital programs for this population. Our findings offer a first step by providing this
design guidance based on the perspectives of prostate cancer survivors.

Our findings indicate that prostate cancer survivors express an interest in overcoming exercise barriers and improving
their PA. When designing any digital program for this group it is valuable to consider features that reduce barriers and
enhance internal motivators (e.g., fun, positive thinking) and offer connections with external motivators, such as
providers, family, and peers. Even though the participants indicated moderate interest using technology to improve
PA in surveys they also provided a range of technology based magic wand ideas to help them walk more (e.g.,
reminders, virtual community etc.). When designing interventions for prostate cancer survivors, care should be taken
to address low technology use and guarded technology interests. Consider utilizing technology that is easy to use and
simple for participants to perform desired functions (e.g., tracking data, setting reminders).

Based on our findings we recommend that a digital walking program for prostate cancer survivors be designed to: 1)
provide well curated and specific content - both authoritative material from professionals and experiential material
from peers; 2) provide individualized feedback (e.g., how one is doing relative to one’s own goals) based on tracked
data or advice from a trusted, experienced source; 3) provide capabilities for a moderated discussion and the ability to
facilitate discussions with peers (e.g., “Has anyone else dealt with the problem I am experiencing?”); and 4) enable
users to work with peers to share interests, motivate each other, and find walking partners rather than compete. Our
results indicate that a well-designed digital walking program that combines fitness trackers with social features (e.g
find walking partners, motivate each other) could potentially increase access to and extend the reach of PA programs
that better support survivors.

Facilitating behavior change for PA is multifaceted and complex. The behavior change intervention literature indicates
that goal setting, feedback, self-monitoring, and social support (i.e. a combination of personal, interpersonal and
environmental factors) should be considered when designing digital behavior change interventions for older adults.24
Yet care needs to be taken to match the behavior change technique to design features that users prefer and most likely
to use. For example, our results indicate that participants lacked interest in competing for badges, even though these
types of digital rewards are a feature that has been shown as beneficial in young adult cancer survivors.24, 25 Our
findings offer a first step by offering guidance for design of engaging digital programs for prostate cancer survivors.

Findings should be evaluated in the context of some limitations. We conducted 3 focus groups in a single metropolitan
area that limit could generalizability of findings. We explored the potential of designing engaging digital technology
to promote PA in a sample that was low in both social media use and fitness tracker use. Lack of experience may have
limited the ability for participants to imagine using these features. In addition, limitations like having certain dominant
individuals in a group and other response biases that are inherent in focus groups might have influenced the data.
Future work should extend the study to a broader audience, such as prostate cancer survivors who do not have access
to support groups, diverse patients with respect to education, race ethnicity, and include more caregivers and friends.
Future research should also evaluate feasibility and effectiveness of a digital walking program designed for prostate
cancer survivors. Despite these limitations, our study has several strengths that include directly engaging future
technology users and our mixed methods approach. Conducting focus groups at existing support group meetings,
where participants regularly meet and are socially connected, facilitated a rich conversation and sharing of ideas.

Conclusion

Given the need to increase PA in prostate cancer survivors but lack of guidance for well-designed digital PA programs
to facilitate behavior change, our findings characterize needs and design preferences for digital walking programs.
Participants expressed interest in improving PA and expressed preferences for design of specific features for program
content, PA feedback, discussion, and teamwork. Digital walking programs that are designed to meet these needs and
preferences could enhance engagement by facilitating an accessible, convenient, and scalable solution for improving
PA among prostate cancer survivors.

References

1. American Cancer Society [Internet]. Key Statistics for Prostate Cancer. Atlanta, GA: American Cancer Society,
February/2021.


22. Dedoose 8.3.43 Web application for managing, analyzing, and presenting qualitative and mixed method research data. SocioCultural Research Consultants, LLC; 2019. www.dedoose.com


A Fusion NLP Model for the Inference of Standardized Thyroid Nodule Malignancy Scores from Radiology Report Text

Thiago Santos, MS,1,3, Omar N. Kallas, MD,2, Janice Newsome, MD,2, Daniel Rubin, MD., MS,4, Judy Wawira Gichoya, MD MS,2,3, Imon Banerjee, Ph.D,2,3,
1Department of Computer Science, Emory University, Atlanta, GA, USA; 2Department of Radiology, Emory School of Medicine, Atlanta, GA, USA; 3Department of Biomedical Informatics, Emory School of Medicine, Atlanta, GA, USA; 4Department of Biomedical Data Science, Stanford University School of Medicine, Palo Alto, CA, USA;

Abstract

Radiology reports are a rich resource for advancing deep learning applications for medical images, facilitating the generation of large-scale annotated image databases. Although the ambiguity and subtlety of natural language poses a significant challenge to information extraction from radiology reports. Thyroid Imaging Reporting and Data Systems (TI-RADS) has been proposed as a system to standardize ultrasound imaging reports for thyroid cancer screening and diagnosis, through the implementation of structured templates and a standardized thyroid nodule malignancy risk scoring system; however there remains significant variation in radiologist practice when it comes to diagnostic thyroid ultrasound interpretation and reporting. In this work, we propose a computerized approach using a contextual embedding and fusion strategy for the large-scale inference of TI-RADS final assessment categories from narrative ultrasound (US) reports. The proposed model has achieved high accuracy on an internal data set, and high performance scores on an external validation dataset.

1 Introduction

Thyroid nodules are very common, with a reported prevalence of 33% to 68% in the healthy population.1 Despite the high prevalence of thyroid nodules, only 1.2-12% are malignant. Diagnostic thyroid ultrasound is critical for nodule detection and characterization. Diagnostic thyroid ultrasounds are used to determine whether a thyroid nodule has features suspicious for malignancy, in which case an ultrasound-guided fine needle aspiration or surgical thyroidecstasy may be indicated for further work-up and management. The Thyroid Imaging Reporting and Data System (TI-RADS) and the American Thyroid Association (ATA), released by the American College of Radiology (ACR), are two existing reporting systems that standardize reporting lexicon, risk stratification, and management guidelines for thyroid nodules. There is significant variation in radiologist practice when it comes to diagnostic thyroid ultrasound interpretation and reporting, due to the existence of these two risk stratification systems, and the frequent use of narrative non-structured reporting. The TI-RADS system, more recently developed, categorizes thyroid nodules with one of five labels based on malignancy risk: TR1-benign, TR2-not suspicious, TR3-mildly suspicious, TR4-moderately suspicious, TR5-highly suspicious; guiding nodule management.

Our hypothesis is that an artificial intelligence (AI) system can be developed to automate the assessment of thyroid nodule malignancy risk from thyroid ultrasound images. To achieve this goal, the machine-learning system would need to be developed to label large numbers of longitudinal imaging examinations using the standardized terminology of TI-RADS reporting system. Subsequently, these annotated exams would be used to train an AI algorithm that can extract thyroid nodule risk directly from the US images, and display risk scores to the interpreting radiologists for verification. Such a system has the potential to reduce practice variation between the radiologists in assessing thyroid nodule risk from US images, streamline imaging workflow, and improve billing processes, through automated TI-RADS coding of the radiology reports. A standardized thyroid nodule database may also help population health and nodule surveillance, and be utilized for secondary analysis including subpopulation evaluation.

Text classification is an important part of Natural Language Processing (NLP). The underlying goal of text classification is to associate a given text string with one or more categories. This association is based on different characteristics of the text string (content, terms, context), under predefined classification taxonomy. Effective text feature selection is one of the biggest challenges faced in NLP and it is a crucial step towards designing a NLP algorithm with better efficiency and performance. The information extraction and text classification aspects of NLP can play a key role in
improving workflow efficiency in medical institutions.

The NLP-based automated inference of TI-RADS category from semi-structured radiology reports is an important first step towards developing an automated AI system to extract thyroid nodule malignancy risk directly from US imaging studies. The lack of standardized radiology reports poses a challenge for NLP-based algorithms. Furthermore, the diversity of the medical language space combined with its very specific terminology increases the difficulty of NLP-based medical text classification. For example, when describing a TI-RADS nodule scored as 3 (mildly suspicious), many domain-specific terms can be used, like: solid, hyperechoic, hypoechoic, isoechoic, and many others. These different TI-RADS descriptors can also be seen with different TI-RADS categories, as the TI-RADS system takes into account the malignancy risk of multiple different imaging characteristics and sums them up.

Several studies have explored the use of machine learning and NLP-based methodology for information extraction, or more specifically, for automated score-based classification of medical reports.\textsuperscript{2-7} Dorothy et. al.\textsuperscript{8} utilized NLP to extract BI-RADS final assessment categories from radiology reports. Sergio et. al.\textsuperscript{9} developed a machine-learning pipeline, using Support Vector Machine (SVM), for the automated extraction of Breast imaging-reporting and data system (BI-RADS) categories from breast radiology reports. Banerjee et. al.\textsuperscript{10} utilized machine learning to infer Liver Imaging Reporting and Data System (LI-RADS) scores, a malignancy risk-stratification system for liver lesions, from structured and unstructured reports through NLP-based transfer learning based on an established LI-RADS ultrasound (US) report database. With the rapid development of deep learning, a variety of novel NLP models and methods have emerged with improved performance, including text classification methods based on word vector models. As examples, Kim et al\textsuperscript{11} applied convolutional neural networks (CNN)\textsuperscript{12} for sentence classification tasks. Lee et al\textsuperscript{13} developed an NLP pipeline by combining recurrent neural networks (RNN) and convolutional neural networks (CNN) to classify short text strings. Zhou et al\textsuperscript{14} interpolated the idea of 2-dimensional maximum pooling operation to create a bidirectional long-term and short-term memory network (BILSTM), enabling the extraction of text features in the temporal and spatial dimensions during a downstream task.

In contrast to the previous approaches, recent advances in NLP techniques can be leveraged for better text classification by exploiting Transformers Encoder-Decoder\textsuperscript{15} word representations. The Bidirectional Encoder Representations from Transformers (BERT) model\textsuperscript{16} proposed by Google utilizes a self-attention mechanism transformer,\textsuperscript{15} and is a context-preserving NLP technique. BERT models outperformed previously published NLP models like word2vec,\textsuperscript{17,18} which were based on context-independent embeddings. This is because BERT can represent words or sequences in a way that captures the contextual information, causing the same sequence of words to have different representations when they appear in different contexts. However, most of the published work in BERT-based text classification in medicine uses only single source text data, while often multi-source data integration is needed for comprehensive representation of a single data point.

In this work, we propose a semi-supervised NLP pipeline for TI-RADS reports that combines BERT contextualized representations with patient demographic information as well as comorbidities, in order to create an ensemble late fusion classification model.\textsuperscript{19} The concatenation of these vectors representations are utilized in a machine-learning setting for inferring TI-RADS of all nodules from semi-unstructured reports based on the five TI-RADS final assessment categories. For this study, we experimented with a set of text featurization methods and multiple traditional machine learning and deep learning models, and compared their performance on three independent datasets.

2 Methods

Figure 1 presents the core processing blocks of the proposed pipeline. The pipeline combines the ‘FINDINGS’ documented in the radiology reports with the ‘CLINICAL INDICATION’ (Past nodule history information from patient) and ‘DEMOGRAPHIC INFORMATION’ (Age, Gender, Race) of the patient to construct an ensemble late fusion classification model.\textsuperscript{19} This fusion model mimics the radiologist’s workflow for scoring a study by adding the necessary additional information. Each processing block is described in the following section.
In order to perform the study and validate the model and results, we collected the following three non-overlapping corpora of ultrasound radiology reports of thyroid nodule screening from two academic healthcare institutions.

Corpus 1. Semi-Structured reports from EUH - With the approval of Emory University Institutional Review Board (IRB), a total of 665 thyroid cancer screening ultrasounds(US) reports from unique patients were extracted from the clinical data warehouse of Emory University Hospital (EUH) between July 2017 and March 2020. The mean age of the sample is 59(+/-14), 81% female, 48% white, 32% African American, 10% Asian, and 10% Others. These reports are written following a strict template, and contain standard sections such as: "HISTORY", "CLINICAL INDICATION", "FINDINGS", and "IMPRESSIONS". Although these reports can be written in a structured and semi-structured format, in this work we are interested on working with semi-unstructured reports, since extraction of scores from structured TI-RADS formatted reports is a trivial task. However, extraction of TI-RADS scores from semi-unstructured reports is challenging since they are documented in free-text and radiologists do not consult any particular vocabulary or taxonomy when describing and scoring each nodule. For this study, 627 of the 665 reports were categorized as semi-structured.

Corpus 2. Prospective Semi-Structured reports from EUH - Similar to corpus 1, a total of 290 thyroid cancer screening ultrasounds(US) reports from unique patients were extracted from the clinical data warehouse of Emory University Hospital (EUH) from April 2020 to October 2020. The mean age of the sample is 59(+/-14), 78% female, 49% white, 38% African American, 5% Asian, and 8% Others. The purpose of collecting a different corpus, from
the most up to date studies available to us, is to test the generalizability of the proposed model. Following the same criteria as corpus 1, we were able to identify 221 out of 290 as semi-unstructured reports.

**Corpus 3. Semi-Structured reports from SHC** - In order to test the generalizability of our algorithm on an external dataset, we have collected over 177 semi-unstructured TI-RADS radiology reports in collaboration with Stanford Healthcare (SHC) acquired between Jan 2015 - Dec 2017. The mean age of the sample is 41(±12), 58% female, 52% white, 25% Asian, and 3% African American. Similar to the EUH dataset, we parsed the reports and extracted the TI-RADS score assigned by the original reader, and after extraction we dropped the code from the text for evaluation of our NLP algorithm.

### 2.2 Report splitter

A TI-RADS report may contain information about multiple nodules. According to the ACR recommendation, radiologists should add detail characteristics of each nodule found, e.g., size, shape, composition, margin, echogenicity. In order to separate each nodule and its corresponding characteristics, we developed a python-based segmentation algorithm for the Finding section of the radiology report using regular expressions to identify key words indicative of a nodule, while maintaining dependencies between anatomical entities. Finally, we combined multiple sentences to formulate the description of nodules. In order to stop label-leakage, we excluded the Impression section of the reports since the final TI-RADS score is often reported explicitly in the impression. The Findings section includes only the characteristics of the nodule abnormalities and therefore, it does not leak any information regarding the TI-RADS final assessment classes.

In order to evaluate the performance of our segmentation algorithm, we manually segmented 50 random reports from Corpus 1. A total of 87 TI-RADS nodules were extracted, with label distribution of: TR1: 13, TR2: 17, TR3: 28, TR4: 21, TR5: 8. Our segmentation was able to identify and extract all 87 nodules correctly. When dealing with free-text reports, the full nodule description can be hard to identify since a nodule can be described in different sentences and paragraphs. Our algorithm was able to correctly extract 82 nodules and their full description, achieving an accuracy of 94%. The partial text description extracted from the 6 remaining nodules were considered satisfactory to identify the TI-RADS category of the nodule.

Once we segmented each EUH report, we identified 1,006 TI-RADS nodules from Corpus 1 and 597 TI-RADS nodules from Corpus 2. A total of 450 unique words were found, with nodule descriptors ranging between 3 to 45 words, with a mean report size of 15 and standard deviation of 6. As for SHC dataset, we were able to extract 177 nodules and their corresponding descriptors. A total of 390 unique words were presented and report size ranged from 4 to 60 words with a mean of 19 and standard deviation of 10. Figure 2 shows the distribution of TI-RADS labels from the EUH (Corpus 1 & Corpus 2) and SHC(Corpus 3) dataset. Given that our study includes the US screening population, the distribution is highly uneven, with very few representative reports with TI-RADS 5 (highly suspicious) and a high number of reports with TI-RADS score of 3 and 4 (mild to moderate suspicion). In addition to ‘FINDING’, we segmented and extracted the ‘CLINICAL INDICATION’ section from the radiology reports when present in the text report.

### 2.3 Text cleaning & preprocessing

We use pre-processing when constructing out baseline models, to reduce the feature-space and get better data generalization. However, the following steps were not applied with Bidirectional Encoder Representations from Transformers (BERT) models. Data pre-processing has been shown to improve the quality of extracted text features. In this work, we transformed the segmented data through a series of standard pre-processing techniques. This was accomplished by applying a conversion to lowercase, removal of stop words (e.g. ‘a’, ‘an’, ‘are’, etc.), removal of low-frequency (< 50) words, and removal of unwanted terms and phrases. We used the Natural Language Tool Kit (NLTK) library to obtain the list of stop words. We discarded identifying details of radiologist, clinicians, patients, and TI-RADS scores from the reports. While we applied a stemming and lemmatization technique for most of the baseline models, only lemmatization was applied to the original version of words when dealing with machine and deep learning models. This is because such models involve some kind of word textual embedding, and we are dealing with pre-trained embeddings, which were trained with original version of the words.
In addition to the usual standard text pre-processing, we also applied some specific pre-processing techniques for radiology domain. When dealing with radiology reports, there are some crucial information embedded in numbers and symbols, such as: measurements, percentage, shape, and others. Thus, it’s crucial to our system to be able to read and understand such information since those are critical information for obtaining the final scoring. We converted all numbers into character strings, e.g., ‘22’ is then mapped to ‘twenty two’. Furthermore, we also preserved all mathematical symbols, like: ‘>’, ‘%’, by converting them to character strings as well, e.g., ‘greater than’ and ‘percentage’.

2.4 Text Featurization and Discriminative models

Baseline Models: (1) BoW/tf-idf + Logistic Regression; (2) BoW/tf-idf + Random Forest; (3) BoW/tf-idf + XGBoost - We designed our baseline models with traditional NLP text representation, i.e., bag-of-word and tf-idf. We experimented with Logistic Regression, Random Forest and Extreme Gradient Boosting based classifiers with feature representation to form a comparative baseline for our deep learning based TI-RADS score prediction models.

Semantic Embedding and Classifier: (4) word2vec pre-trained/self-trained + Logistic Regression; (5) word2vec pre-trained/self-trained + Random Forest; (6) word2vec pre-trained/self-trained + XGBoost - For a more comparative text featurization approach, we applied Word2Vec which is a semantic embedding based neural network which generates a non-sparse numeric representation of the documents. This representation partially alleviates the problem of limited vocabulary faced by traditional bag-of-words and tf-idf models which simply ignore any relevant out-of-vocabulary words and preserve the semantics of the words while generating the numeric representations. To generate these embeddings representation, we computed the average corresponding vectors of all nodule description words. In our analysis, we experimented word2vec model with and without transfer learning strategy, which we call pre-trained and self-trained word2vec, respectively. For transfer learning, a word2vec model trained on the Google News dataset is fine-tuned with EUH radiology reports. We also experimented training our own version of word2vec over EUH reports. We used the skip-gram model with a vector length of 300, a window width of 5, and default settings for all other parameters. No vectors were built for terms occurring fewer than 6 times in the corpus.

Sequential and Convolutional Deep Learning Models: (6) LSTM (random initialization); (7) LSTM (word2vec initialization); (8) 1DCNN (random initialization); and (9) 1DCNN (word2vec initialization) - We experimented with sequential models - Long short-term memory (LSTM) and One-dimensional convolutional neural networks (1D-CNN). LSTM and 1D-CNN models usually include an Embedding layer as the first layer of the model. The objective of this layer is to learn a vector representation of each word in the training vocabulary set. In both approaches, we have experimented with two different variations, i) random initialization and ii) initialization with word2vec model weights. In the second approach, instead of initializing the model with random weights, we take the word2vec representation of the word as an input. This approach enables the model to take advantage of unlabeled data present in the form of unstructured radiology reports and learn semantic context beforehand. At the same time, the model is able to fine-tune the word-embedding layer for the particular language space while training the prediction model for TI-RADS score prediction. To address the variation in the size of the radiology reports, we used padding with a maximum report length of 30 words.

BERT Ensemble Classification Model - One of the biggest challenges with employing word2vec and tf-idf for radiology report parsing is how to handle unknown, out-of-vocabulary (OOV), and similar words (abbreviations, acronyms). In radiology reports, the use of synonyms and related words are widely used, depending on the preferred style of radiologist. This is a challenge because particular words may be used infrequently in a large radiology report corpus, even when it is collected from the same institution. BERT architectures have been devised to consider the context for each occurrence of a given word, providing a sub-word contextualized word embedding representation that will be different according to the sentence and its meaning. We are using BERT-Base Uncased model with a fully connected layer on top to fine-tune the whole model to a TI-RADS downstream classification task. Our implementation is built using the PyTorchTransformers library by huggingface. We trained the BERT model using an AdamW optimiser with a learning rate set to $10^{-5}$. For each experiment set, the training was run with an early-stopping patience of 100 epochs. The optimal model was then used to produce the reported results. The maximum
Figure 3: Fusion-BERT Classification System. The output probabilities of the three independent models and concatenated and fed into a logistic regression meta-learner.

sequence length was set at 128 and the batch size at 32. The model architecture and training phase are illustrated on Figure 3. In our preprocessing, a “[CLS]” symbol is added before input texts. It is used by the transformer to extract features from texts and encode global information. The output of highest hidden layer at the “[CLS]” position is taken as a sentence-level feature. Subsequently, a fully connected layer is used to output text classification probability values.

2.5 Fusion Model

Our goal is to build a supervised fusion model classifier to automatically recognize the TI-RADS categorization of the nodule. First, we trained three independent models: (1) nodule data classifier: takes a vector representation of the nodule data as input; We select the vectorization scheme that performed optimally for text report classification. (2) EHR classifier: Takes a concatenation of vectors representation as one-hot-encoding of the demographic information of the patient (Age, Gender(Male/Female), and Race (African American, Asian, White, Others, Unknown); (3) Clinical Indication classifier: Takes a vector representation as tf-idf of the clinical indication(history of patient) as input. BERT\textsuperscript{16} performed optimally over all the text vectorization models and was used to build the nodule data classifier, while Random Forest classifier was applied on EHR and Clinical Indication. Next, we create a late fusion-model that concatenate the output probabilities of the 3 models: BERT + nodule description; Random Forest + patient demographic information; and Random Forest + patient clinical indication. Finally, we add a meta-learner as a last layer, a standard non-parametric Logistic Regression classifier in its default configurations (stochastic average gradient solver, intercept scaling = 1, $l_2$ penalty). The reason behind adding a meta-learner is to allow the fusion-model to learn the relevance of each independent model for TI-RADS score prediction. The fusion-model is illustrated in figure 3.

2.6 Statistical Validation

We collected over 848 semi-structured EUH radiology reports (Corpus 1 and Corpus 2) for which we were able to segment 1006 and 597 nodules for our retrospective dataset and prospective dataset respectively (see Sec. 2.1, 2.2). We leveraged this set of nodules to train, test and validate our model, by randomly splitting our corpus 1 into 805 (80%) for training and 201 (20%) for testing purpose. In our second set of experiments, we expanded the scope of the work to include external validation with Stanford SHC dataset (corpus 3). We were able to extract and segment 177 nodules from SHC radiology reports. To evaluate our model, we applied four standard statistical metrics commonly used to validate NLP inference and for deep learning classification tasks. They are Accuracy, Precision, Recall and F1 score. F1 is the standard metric for this task; it is calculated as the harmonic average of the precision and recall. We also reported two addition evaluation measures, 1-margin Accuracy and 1-margin F1 score. Both measures consider a prediction to be correct if it differs from its ground truth by a margin of less than or
equal to 1. Such predictions are useful for clinical purposes because the descriptors for adjacent TI-RADS categories are often overlapping (e.g., TR 3 and TR 4, mild and moderate malignancy). Moreover, while agreement for the indication to biopsy was substantial to near-perfect, lower interobserver agreement has been observed between readers for neighbouring TI-RADS categories.

### 2.7 Interface Design

We developed a user-friendly graphical user interface (GUI) for our system to upload data and obtain prediction results in batch. The interface can handle a single nodule classification as well as prediction over multiple reports stored in a single file. For a single nodule prediction, the user can input a nodule description and the application will predict the output with the highest probability. In this module, the application also gives the probability of all categories, giving a better indication and explanation to the user. The user can also load a complete radiology report and perform classification on all nodules at once. Once the file is loaded, our application automatically segments every report to extract all nodules. Figure 4 illustrates an example of our Interface when loading a file with US radiology reports. Figure 4(a), and also with a single nodule description in Figure 4(b).

### 3 Experiments and Results

**Internal hold-out test data:** Simple baseline models like tf-idf performed very poorly as the thyroid ultrasound report dataset had a complex vocabulary and a very sparse language space. Traditional machine learning models combined with neural network semantic embeddings, like word2vec, had the best baseline performance for TI-RADS score prediction. Pre-trained word2vec models are trained using a large generic text corpus (like the Google News dataset), and leverage optimal embeddings when fine-tuning a language space for medical clinical notes through transfer learning. Given the highly specialized text domain of thyroid ultrasound reports, the combination of self-trained word2vec which was trained from scratch on a large amount of EUH radiology thyroid ultrasound data and Random Forest achieved a F1 score of 0.875, an increase of 1% when compared with the pre-trained word2vec model, becoming a hard baseline to outperform. However, BERT achieved an even greater F1 performance of 0.96 on Corpus 1 hold-out test set, with an improvement of 11.5% when comparing with self-trained word2vec model + Random Forest. The proposed pipeline, an Ensemble Fusion-BERT, was able to outperform BERT itself, showing an improvement of 2%. This illustrates how our model can mimic the radiologist’s workflow of scoring a TI-RADS nodule by adding the necessary additional information if needed. Table 2 summarizes class-wise Precision, Recall, F1-score, and 1-Margin metric of the proposed Ensemble Fusion-BERT model. Table 2 also illustrates the performance of the Fusion-BERT model on each class label for the: i) Corpus 1 - Test set; ii) Corpus 2 - Prospective Validation; and iii) Corpus 3 - SHC external validation. 1-margin metrics are also included. In addition, to present the class-level comparative performance in a more interpretable manner, the corresponding confusion matrices are displayed in graphical form in Figure 5. The numeric values in the confusion matrices show how many nodules were correctly classified as well as
how many were misclassified.

**Internal Prospective Validation:** 597 thyroid nodule description modules of the most recent thyroid US reports from April 2020 to October 2020 (see Sec. 2.1) were used to test the generalizability of the model. Our hypotheses is that most recent reports use different template structures and lexicon, which may result in an overall lower performance, since our model was trained on report data starting from 2017, prior to the implementation of the TI-RADS scoring system and template at our institution. While our model was able to achieve a F1 score of 0.97 on the Corpus 1 test set, it achieved a slightly lower F1 score of 0.89 on Corpus 2 (Prospective Validation Set). However, when using 1-Margin metrics the performance of our model on both, the test and validation sets, yield a F1 score of 0.99. This suggests that even though our model had a slight drop of performance on the prospective data, it is still applicable for curating a large-scale labeled imaging dataset with acceptable accuracy.

**External validation (EV) - SHC Dataset:** New participant level data from SHC, external to those used for model development and testing, were used to assess the reliability of our model’s predictions at a different institution. This is a very important step to validate a model to a potentially non-generalizable problem, where the model can only perform well on the same dataset it was trained. As seen from the classification results in Table 2, our proposed model demonstrated successful generalization on the Stanford SHC radiology reports (Corpus 3) with an average F1 score of 0.85, and 1-Margin F1 score of 0.94.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Feature Representation</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>bag of words</td>
<td>0.612</td>
</tr>
<tr>
<td></td>
<td>tf-idf</td>
<td>0.660</td>
</tr>
<tr>
<td></td>
<td>pre-trained word2vec</td>
<td>0.643</td>
</tr>
<tr>
<td></td>
<td>self-trained word2vec</td>
<td>0.682</td>
</tr>
<tr>
<td>Random Forest</td>
<td>bag of words</td>
<td>0.74</td>
</tr>
<tr>
<td></td>
<td>tf-idf</td>
<td>0.759</td>
</tr>
<tr>
<td></td>
<td>pre-trained word2vec</td>
<td>0.862</td>
</tr>
<tr>
<td></td>
<td>self-trained word2vec</td>
<td>0.875</td>
</tr>
<tr>
<td>XGBoost</td>
<td>bag of words</td>
<td>0.741</td>
</tr>
<tr>
<td></td>
<td>tf-idf</td>
<td>0.775</td>
</tr>
<tr>
<td></td>
<td>pre-trained word2vec</td>
<td>0.862</td>
</tr>
<tr>
<td></td>
<td>self-trained word2vec</td>
<td>0.851</td>
</tr>
<tr>
<td>LSTM</td>
<td>LSTM embedding layer</td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td>word2vec embedding layer</td>
<td>0.77</td>
</tr>
<tr>
<td>1DCNN</td>
<td>CNN embedding layer</td>
<td>0.71</td>
</tr>
<tr>
<td></td>
<td>word2vec embedding layer</td>
<td>0.79</td>
</tr>
<tr>
<td>BERT</td>
<td>BERT</td>
<td>0.96</td>
</tr>
<tr>
<td>Ensemble Fusion-BERT</td>
<td>BERT + one-hot + tf-idf</td>
<td><strong>0.98</strong></td>
</tr>
</tbody>
</table>

**Table 1:** TI-RADS inference table results over randomly hold-out 201 nodules of Corpus 1 test set.

<table>
<thead>
<tr>
<th>Class Label</th>
<th>Corpus 1</th>
<th>Corpus 2</th>
<th>Corpus 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
<td>F1</td>
</tr>
<tr>
<td>1</td>
<td>0.86</td>
<td>0.93</td>
<td>0.93</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0.93</td>
<td>0.96</td>
</tr>
<tr>
<td>3</td>
<td>0.98</td>
<td>0.99</td>
<td>0.98</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0.92</td>
<td>0.96</td>
</tr>
<tr>
<td>Wgt. Avg. Ensemble Fusion-BERT</td>
<td><strong>0.98</strong></td>
<td><strong>0.98</strong></td>
<td><strong>0.98</strong></td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.98</td>
<td>0.89</td>
<td>0.85</td>
</tr>
<tr>
<td>Confidential Interval</td>
<td>95%</td>
<td>95%</td>
<td>95%</td>
</tr>
<tr>
<td>1-Margin Accuracy</td>
<td>0.99</td>
<td>0.99</td>
<td>0.94</td>
</tr>
<tr>
<td>1-Margin F1 Score</td>
<td><strong>0.99</strong></td>
<td><strong>0.99</strong></td>
<td><strong>0.94</strong></td>
</tr>
</tbody>
</table>

**Table 2:** TI-RADS score inference table results from Ensemble Fusion-BERT performance on: i) Corpus 1 - EUH Test set, ii) Corpus 2 - EUH Prospective set, and iii) Corpus 3 - Stanford SHC external validation set.
4 Conclusion

In this work, we experimented with a wide variety of natural language processing and text mining techniques to process the text of thyroid US reports so that machine-learning techniques can be applied to them for TI-RADS score inference. We propose an effective late fusion model for thyroid score extraction which tackles the complex radiology language space using a bi-directional transformer model, which takes into account relevant patient clinical history and demographics. We evaluated our model by testing its performance over a prospective and external cohort. Being trained on a single institution’s data, our model performed exceptionally well for our own, and outside institution, data; indicative of our model’s generalizability. Hence we believe that our model will be able to effectively extract nodule TI-RADS scores from unstructured and semi-structured radiology reports. We plan to deploy the graphical application powered by the Ensemble Fusion BERT developed in this project to standardise point of care thyroid ultrasounds being performed by non radiologists in the endocrine clinic.

References


Predictors of Retention for Community-Based Telehealth Programs: A Study of the Telehealth Intervention Program for Seniors (TIPS)

Melody K Schiaffino, PhD, MPH1, Zhan Zhang, PhD2, David Sachs, EdD2,3, John Migliaccio, PhD3, Jina Huh-Yoo, MHCI, PhD4

1School of Public Health, San Diego State University, San Diego, CA; 2School of Computer Science and Information Systems, Pace University, New York, NY; 3Westchester Public/Private Partnership for Aging Services, White Plains, NY; 4College of Computing and Informatics, Drexel University, Philadelphia, PA

Abstract
Community-based telehealth programs (CTPs) allow patients to regularly monitor health at community-based facilities. Evidence from community-based telehealth programs is scarce. In this paper, we assess factors of retention—patients remaining active participants—in a CTP called the Telehealth Intervention Programs for Seniors (TIPS). We analyzed 5-years of data on social, demographic, and multiple chronic conditions among participants from 17 sites (N=1878). We modeled a stratified multivariable logistic regression to test the association between self-reported demographic factors, caregiver presence, status of multiple chronic conditions, and TIPS retention status by limited English proficient (LEP) status. Overall, 59.5% of participants (mean age: 75.8yrs, median 77yrs, SD 13.43) remained active. Significantly higher odds of retention were observed among LEP females, English-speaking diabetics, and English proficient (EP) participants without a caregiver. We discuss the impact of CTPs in the community, the role of caregiving, and recommendations for how to retain successfully recruited non-English speaking participants.

Introduction
Increasing demand for healthcare in response to the needs of a growing aging population, and more recently to the rapid adoption of telemedicine due to the COVID-19 (SARS-CoV-2) pandemic, has expedited the urgency to ensure telehealth that is safe, effective, efficient, patient-centered, timely and equitable. Access to high-quality care for aging populations is increasingly complex due to the unique needs of older adults as their diversity in race/ethnicity, socioeconomics, and complex health status demonstrates. This group accounts for the largest population affected by multiple chronic conditions (MCC) or having two or more chronic diseases. While 25% of all Americans live with an MCC, this figure increases to over 60% among older adults.1 Treating MCCs in fragmented healthcare delivery systems creates higher costs overall, and results in poorer quality for older Americans. This risk is compounded for racially and ethnically diverse older adults and those who speak English less than “very well.” They are considered limited English proficient (LEP) because they are more likely to experience care delays and communication barriers since their primary language is not English. Further, as older adults, they are also facing higher rates of cognitive impairment, dementia, reduced functional status and increased emergency visits among Medicare beneficiaries.2 Evidence underscoring the benefits of aging-in-place3 for older adults who wish remain in their homes include better mobility, cognition, social status, and lower depression risk.4 In response, innovations in telehealth and telemedicine interventions have become ubiquitous.5, 6 This technology has enabled the delivery of safe and effective care for Americans who need to access care remotely or require constant monitoring.9,11 Many telehealth technologies for older adults are home-based. Despite offering many benefits, home-based telehealth programs face great challenges in adoption and uptake due in part to the usability and cost issues of telehealth device.12, 13 In recent years, community-based telehealth (CTPs) programs—offering telehealth services in community settings (e.g., congregate housing, community centers)—are gaining momentum, because 1) they are less equipment intensive, thus less cost is transferred to the patients, and 2) in-person support is available for immediate monitoring and feedback.9,14,15 However, issues persist among older Americans with complex needs that impede successful uptake related to the usability, tailoring, and feasibility make technology enabled healthcare inaccessible for many older adults and can be particularly prohibitive for low-income, diverse older adults who are LEP.16, 17 This is in addition to the already significant barriers related to accessing healthcare due to language and communication.18, 19
Given the scarce evidence on CTPs, we analyzed 5-years of data on social, demographic, and multiple chronic conditions among participants in the community-based Telehealth Intervention Program for Seniors (TIPS) to provide insights to the feasibility of CTP implementation, particularly on retention. By examining TIPS program retention in a diverse, low-income population of older adults, our study contributes to understanding the roles of language
proficiency, multiple chronic conditions status, and the role of caregivers in retaining participants in CTPs and how informatics approach can address identified challenges.

**Background**

Telehealth is essential for monitoring high-risk aging populations with chronic conditions in the community in order to ensure timely and high-quality care, reduced care utilization and hospital readmissions, and improved outcomes.\(^{20, 21}\) There are different types of telehealth programs designed to provide care to older adults. The most commonly and widely-adopted model is home-based telehealth interventions. Numerous studies have demonstrated the effectiveness of home-based telehealth.\(^{7, 8, 10, 11, 22}\) Home-based telehealth is associated with improved self-management of multiple chronic conditions (MCC) among older adults,\(^{23, 24}\) and may benefit those that experience mobility and transportation barriers most.\(^{25}\) Seminal work has focused on examining different aspects of this type of telehealth program, such as user acceptance,\(^{12, 13}\) usability,\(^{26}\) and effectiveness.\(^{10}\) However, this approach has several barriers, hindering its uptake and adoption. For example, the usability of home-based telehealth systems for the aging population is problematic as many older adults may not be able to use and navigate the system as well as their younger counterparts.\(^{27}\) Even more concerning is that current home-based telehealth technology requires installation of equipment in a user’s home, regular maintenance, and individualized healthcare services, all of which pose significant financial burden on older adults.\(^{28}\) Thus, home-based telehealth services are often not an optimal option for many older adults and less so for those with lower socioeconomic status.

To deliver remote monitoring to large-scale aging populations at a lower cost, an alternative approach—community-based telehealth programs (CTPs)—is increasingly being deployed in recent years.\(^{8, 14, 15}\) Such programs are expected to provide unprecedented opportunities for low-income, high-risk older adults to play an active role in self-management, and in turn, reducing the rates of hospital visits in older adults and the burden of health and social care services.\(^{15}\) A study assessing the benefits of a CTP in a senior-living facility described as “high-intensity” telemedicine found reductions in emergency visits and readmissions.\(^{29}\) Despite the benefits, CTPs face salient challenges and barriers in engaging users because older adults may not necessarily know the potential benefits of these interventions and lack motivation to receive health services through CTPs over time.\(^{30, 31}\) The lack of user engagement and retention with CTPs could negatively affect the effectiveness of delivering and promoting community-based healthcare self-management.

Literature evaluating home-based telehealth for LEP populations with MCC emphasized the need for language access.\(^{32, 33}\) interventions included mobile interpreting apps, and texting-based self-management interventions.\(^{34, 35}\) Our review found a lack of telehealth resources for LEP patients an issue that was highlighted early into the COVID-19 pandemic.\(^{32}\) We found no evidence evaluating CTPs for older adults with LEP. Therefore, it is critical to examine the factors affecting vulnerable user retention in CTPs. Our study will contribute to bridging this research gap. TIPS brings affordable and easier-to-access telehealth to low-income older adults who often face barriers due to age, language, literacy, and costs.

**TIPS Program Overview.** TIPS is a CTP implemented in the U.S. Northeast that provides remote patient monitoring and wrap-around social services to financially vulnerable older adults living in congregate housing or who attend local community centers. Older adults were eligible to enroll in TIPS if they were over 55 years old and registered as a Medicare and/or a Medicaid beneficiary. Older adults access this program by visiting their local community center, or their own long-term care facility that is staffed weekly by trained Telehealth Technician Assistants (TTAs). TTAs are recruited and trained by TIPS to operate the telehealth devices and technologies (e.g., blood pressure cuffs, pulse oximeters, and tablet computers for data entry) and help older adults with onsite assessment of physiological measures (e.g., blood pressure, pulse oximetry, weight, etc.). TTAs also record details about each participant’s evolving medical status (e.g., hospitalization, medication taken, fallen, and overall feeling). Such programs are responsive to the unique needs of older adults, often including social and technology support.

During the initial visit, each participant was asked to fill out an intake questionnaire, which assessed their socioeconomic status (e.g., income, living arrangement, caregiver), demographics (e.g., age, sex, primary language, ethnicity), Medicaid/Medicare or other received benefits, medical history, and multiple chronic conditions. In addition to the questionnaire, TTAs also collected a set of physiological biomarkers (heart rate, blood oxygen saturation, blood pressure, and body weight) and self-reported incidence of (1) ER admission, (2) single hospitalizations, and (3) readmissions less than 30 days following hospital discharge, in the 12 months before enrollment. The intake information and initial monitoring data formed the baseline standards for each participant.

Following initial intake, participants were instructed to visit a TIPS site at least once per week, set-up was the same at every site. At each visit, TTAs used the devices to measure the participant’s physiological biomarkers and record details about their recent medical history using a five-question survey: (1) Have you changed medications since your
last screening visit? (2) Have you changed your medication dosage since your last screening visit? (3) Have you fallen since your last screening visit? (4) Have you been hospitalized or had an ER visit since your last screening visit? (5) How are you feeling today? (response options include “very good”, “good”, “feeling OK”, “feeling a little down”, “not too well”, or “terrible”). The results of the assessment were transmitted to a secure, HIPAA-compliant data server, and then reviewed by a team of TIPS nurses. When a participant was involved in the program, their visit frequency was tracked over time. Individuals who missed more than 4 consecutive weeks of TIPS monitoring were contacted by their designated TIPS nurse regarding why they had stopped using the telehealth service. The reason for the stopped use of TIPS service for each participant was recorded, such as deceased, moved away, or preferred to not continue with the TIPS service. This data indicated whether or not a participant stayed active in the program. Studies using these interventions have demonstrated successful monitoring of older adults with heart failure and MCC. However, a lack of user engagement with such programs has suggested variable implementation of CTPs. To address this problem, we assessed factors associated with older adults staying active in TIPS.

Methods

Dataset. Our study is an observational comparison of baseline characteristics of participants retained and not retained in the TIPS program since it was deployed. Data were collected by staff on-site and transmitted asynchronously (store-and-forward) for review by a team of Registered Nurses, who review the data and contact the participant if their health data triggered an alert during their visit. We merged baseline data from TIPS program sites across New York, New Jersey, Pennsylvania, and Connecticut (N=1878). These data include all baseline data for available sites at the time of sampling (November 2014–October 2019). All analyses and co-authors’ data use agreements were approved by the Pace University IRB.

Variables. We developed a dichotomous outcome variable to represent patients who stayed active with TIPS at the time of analysis to equal “1” and those who were not active equal to “0”. Independent factors included demographics: sex, age, language, caregiver status, MCC, and Medicaid eligibility. LEP participants were identified at enrollment by their self-reported primary language. If any language other than English was selected, they were classified as LEP=“1”. To assess caregiver status, participants were asked if someone (family member or not) took care of them, either living with them or not. If someone did have a caregiver, they were classified as CG=“1”; the reference group for all of these variables was “0”. MCC and other conditions reported at baseline were coded as “1” if yes and “0” if no. MCC and other conditions were self-reported as: hospitalization before TIPS enrollment in the last 12 months, obesity, falls, fractures in addition to self-reported diagnosis of depression, early-stage dementia/Alzheimer’s disease (ADRD), congestive heart failure (CHF), chronic obstructive pulmonary disease (COPD), coronary artery disease (CAD), diabetes mellitus (type 1 or 2), hypoglycemia, liver disease, renal failure, stroke, hypertension, and hypotension. These data were collected at the intake and coded as “1” if selected to indicate yes and “0” if not. We created an additional dummy variable to account for data collection site closures to assess if this had any impact on the sample.

Analysis. We used chi-square analysis to compare group differences across select factors for model selection. We found a significant interaction between caregivers and English proficiency (p=0.0233) and thus presented all descriptive and inferential results stratified by English and non-English proficiency and whether they stayed active in the program (Table 1). Based on our findings and diagnostics, we modeled stratified multivariable logistic regressions to test the association between demographic and MCC factors and the odds of staying active in the TIPS program by language proficiency. All data management and analyses were conducted using SAS 9.4 (Cary, N.C).

Results

Overall, 59.7% of participants stayed active in the TIPS program for the study period. Retention among English speakers (62.6%) was significantly better than that of LEP respondents (47.3%, p<.0001). The average participant age was 75.8 years (median 77.0, SD 13.43), most were female (78%), and did not have a caregiver (95.3%). Also, 36.5% reported having Medicaid and 24.8% were LEP. Table 1 demonstrates comparisons by English proficiency. Over 50% of LEP participants were Medicaid recipients compared with EP (p<.0001), and LEP were significantly more likely to report having a caregiver compared with EP, 6.9% vs. 2.9% (p=0.0004). Both groups reported lacking a caregiver at over 90%.

Table 1. Select characteristics of TIPS program participants by English Proficiency (N=1878).

<table>
<thead>
<tr>
<th>Language Proficiency</th>
<th>English Proficient</th>
<th>LEP</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>n (col%)</td>
<td>n (col%)</td>
<td></td>
</tr>
<tr>
<td>Age Group</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1091
Table 2 compares retention status across EP and LEP participants. We observed that 65 to 85-year-old participants were most likely to report staying active in TIPS compared to other age groups, independent of language. Although not significant, findings suggest a trend among female participants being more likely to stay active in the program compared with males also across both EP and LEP. Retention was higher among EP participants on Medicaid (51.1%, p=0.0048) compared with LEP Medicaid recipients (41.6%, p=0.0096). Among EP participants, retention was significantly higher when no caregiver was reported (62.1%) compared to 38.9% of EP with a caregiver (P=.0048). Conversely, LEP participants did not hold the same pattern. Although not significant, LEP participants reported similar retention with (48.3%) or without (47.3%) a caregiver. Retention of LEP participants without an MCC was higher (50.3%) compared to 39% of LEP with an MCC (p=0.0351). Chronic conditions that were significantly associated with retention status for English participants were depression, history of fracture, hypoglycemia, and hypertension (Table 2). For LEP participants we found only COPD to be significant.

Results from the stratified multivariable model are in Table 3, showing that age was associated with greater odds of retention in TIPS among both EP and LEP participants. Adjusted odds of retention were 71% (AOR 1.71; 95CI 1.12-2.58) greater for English speaking adults age 65-74 and nearly 3-fold greater (AOR 2.60, 95CI 1.21-5.60) for LEP participants in the same age group compared with the reference group (<65 years). English participants age 75-85 reported 65% greater odds (AOR 1.65; 95CI 1.11-2.45), compared with similar 3-fold greater odds for LEP participants (AOR 2.86; 95CI 1.37-5.96)
Table 2. Characteristics of *TIPS* program participants by English Proficiency and Retention Status (N=1878)

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>English Proficient</th>
<th>Limited English Proficient (LEP)</th>
<th>Pvalue&lt;sup&gt;3&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Stayed Active n (row%)</td>
<td>Not Active n (%)</td>
<td></td>
</tr>
<tr>
<td>Age Group</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;65 years</td>
<td>246</td>
<td>119 (48.4)</td>
<td>127(51.6)</td>
</tr>
<tr>
<td>65-74 years</td>
<td>310</td>
<td>222(71.6)</td>
<td>88(28.4)</td>
</tr>
<tr>
<td>75-85 years</td>
<td>395</td>
<td>277(70.1)</td>
<td>118(29.9)</td>
</tr>
<tr>
<td>85+ years</td>
<td>374</td>
<td>212(56.7)</td>
<td>162(43.3)</td>
</tr>
<tr>
<td>Sex</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>1046</td>
<td>661(63.2)</td>
<td>385(36.8)</td>
</tr>
<tr>
<td>Male</td>
<td>279</td>
<td>169(60.6)</td>
<td>110(39.4)</td>
</tr>
<tr>
<td>Medicaid Recipient</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>438</td>
<td>224(51.1)</td>
<td>214(48.9)</td>
</tr>
<tr>
<td>No</td>
<td>887</td>
<td>606(68.3)</td>
<td>281(31.7)</td>
</tr>
<tr>
<td>Caregiver (CG)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CG</td>
<td>36</td>
<td>14(38.9)</td>
<td>22(61.1)</td>
</tr>
<tr>
<td>No CG</td>
<td>1201</td>
<td>746(62.1)</td>
<td>455(37.9)</td>
</tr>
<tr>
<td>Study Site Status</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Site Open</td>
<td>711</td>
<td>404(56.8)</td>
<td>307(43.2)</td>
</tr>
<tr>
<td>Site Closed</td>
<td>614</td>
<td>426(69.4)</td>
<td>188(30.6)</td>
</tr>
<tr>
<td>Chronic Conditions (MCC)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No MCC</td>
<td>920</td>
<td>572(62.2)</td>
<td>348(37.8)</td>
</tr>
<tr>
<td>MCC</td>
<td>405</td>
<td>258(63.7)</td>
<td>147(36.3)</td>
</tr>
<tr>
<td>Depression</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>93</td>
<td>38(40.9)</td>
<td>55(59.1)</td>
</tr>
<tr>
<td>No</td>
<td>1232</td>
<td>792(64.3)</td>
<td>440(35.7)</td>
</tr>
<tr>
<td>Hospitalized&lt;sup&gt;2&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>326</td>
<td>194(59.5)</td>
<td>132(40.5)</td>
</tr>
<tr>
<td>No</td>
<td>999</td>
<td>636(63.7)</td>
<td>363(36.3)</td>
</tr>
<tr>
<td>Falls</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>42</td>
<td>21(50.0)</td>
<td>21(50.0)</td>
</tr>
<tr>
<td>No</td>
<td>1283</td>
<td>809(63.1)</td>
<td>474(36.9)</td>
</tr>
<tr>
<td>Fracture</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>20</td>
<td>6(30.0)</td>
<td>14(70.0)</td>
</tr>
<tr>
<td>No</td>
<td>1305</td>
<td>824(63.1)</td>
<td>481(36.9)</td>
</tr>
<tr>
<td>Dementia/ADRD</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>19</td>
<td>10(52.6)</td>
<td>9(47.4)</td>
</tr>
<tr>
<td>No</td>
<td>1306</td>
<td>820(62.8)</td>
<td>486(37.2)</td>
</tr>
<tr>
<td>CHF</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>45</td>
<td>28(62.2)</td>
<td>17(37.8)</td>
</tr>
<tr>
<td>No</td>
<td>1280</td>
<td>802(62.7)</td>
<td>478(37.3)</td>
</tr>
<tr>
<td>COPD</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>147</td>
<td>94(64.0)</td>
<td>53(36.1)</td>
</tr>
<tr>
<td>No</td>
<td>1178</td>
<td>736(62.5)</td>
<td>442(37.5)</td>
</tr>
</tbody>
</table>

Table 2 (Continued). Characteristics of *TIPS* program participants by English Proficiency and Retention Status (N=1878).
Table 3. Patient Demographic and Self-Reported Factors Associated with Retention in TIPS by English Proficiency Status (N=1878)

<table>
<thead>
<tr>
<th>Condition</th>
<th>English Proficient</th>
<th>Limited English Proficient (LEP)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Stayed Active n (row%)</td>
<td>Not Active n (%)</td>
</tr>
<tr>
<td>Coronary Artery Disease</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>128</td>
<td>89 (69.5)</td>
</tr>
<tr>
<td>No</td>
<td>1197</td>
<td>741 (61.9)</td>
</tr>
<tr>
<td>Diabetic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>282</td>
<td>189 (67.0)</td>
</tr>
<tr>
<td>No</td>
<td>1043</td>
<td>641 (61.5)</td>
</tr>
<tr>
<td>Hypoglycemia</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>19</td>
<td>7 (36.8)</td>
</tr>
<tr>
<td>No</td>
<td>1306</td>
<td>823 (63.0)</td>
</tr>
<tr>
<td>Hypertension</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>679</td>
<td>444 (54.4)</td>
</tr>
<tr>
<td>No</td>
<td>646</td>
<td>386 (59.8)</td>
</tr>
<tr>
<td>Liver Disease</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>17</td>
<td>5 (29.4)</td>
</tr>
<tr>
<td>No</td>
<td>1308</td>
<td>825 (63.0)</td>
</tr>
<tr>
<td>Obesity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>110</td>
<td>71 (64.6)</td>
</tr>
<tr>
<td>No</td>
<td>1215</td>
<td>759 (62.5)</td>
</tr>
<tr>
<td>Hypotension</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>33</td>
<td>24 (72.7)</td>
</tr>
<tr>
<td>No</td>
<td>1292</td>
<td>806 (62.4)</td>
</tr>
<tr>
<td>Renal</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>24</td>
<td>16 (66.7)</td>
</tr>
<tr>
<td>No</td>
<td>1301</td>
<td>814 (62.6)</td>
</tr>
<tr>
<td>Stroke</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>75</td>
<td>46 (61.3)</td>
</tr>
<tr>
<td>No</td>
<td>1250</td>
<td>784 (62.7)</td>
</tr>
</tbody>
</table>

Notes: Some values may not add up to full N=1878 due to missing values and/or rounding. Hospitalized in the last 12 months prior to program participation. If cell value is n=5 or less, Fisher’s Exact Two-sided P-value is reported. MCC= sum(FallHX, ADRD, CHFHX, CorArtD, Diabetic, HypoglycemiaHx, HypertensionHx, LiverHx, ObeseHx, HypotensionHx, renalHx, StrokeHx). No MCC=0; 1 chronic conditions; MCC=2+ chronic conditions

Only LEP participants had persistent significantly greater odds of retention for adults over 85 years (AOR 2.63; 95CI 1.19-5.80). Sex was significantly associated with retention but only among LEP participants with females reporting 71% greater odds than males (AOR 1.71; 95CI 1.06-2.75). English proficient and LEP Medicaid recipients reported similar significantly lower odds of staying active (AOR 0.61 and AOR 0.60, p<0.05). Only English participants were associated with significant 3-fold greater odds of staying active if they report lacking a caregiver (AOR 2.53; 95CI 1.23-5.22).

Discussion

The TIPS program demonstrated efficacy in reducing rehospitalizations. However, retention is key to sustaining such benefits of the program. Our study provides evidence on retention in CTPs, which is lacking in the literature. Compared with available evidence on CTPs, our study population consisted of a higher proportion of participants who were LEP (1 in 4), a hard-to-reach population. Even with this vulnerable population, 47.2% participants stayed active in TIPS throughout the 5-year study period. Such long-term evidence is scarce in the literature, and other similar programs with less vulnerable population had faster rate of attrition (e.g., telehealth kiosks where 47% remained active after only 10 months). Our findings from a diverse and geographically heterogeneous sample of community-based older adults with MCC brings key implications to building knowledge on the predictive factors of retention in CTPs among vulnerable aging populations.
Our study population comprised of a diverse sample of low-income older adults with LEP and chronic conditions, many of whom classified as having MCC. We found that, regardless of their chronic condition status, EP and other social-demographic factors were predictive of greater odds of retention in TIPS. Thus, it is the predisposing factors that participants have little control over (e.g., language, income) that may be a greater contributor to retention than
any physical element. Another factor is the health care delivery systems, as structured, struggles to support self-management change in vulnerable populations due to its inability to target well. CTPs appear to fill a critical need in this respect. It is imperative for participants with MCC to experience benefits from telehealth programs like TIPS. While our findings showed that LEP participants with MCC are less likely to be retained, significantly better odds of retention were observed among English proficient participants with select chronic conditions. Consideration for tailoring the socio-technical design structure of TIPS and other CTPs to language and culturally diverse populations may be an opportunity to retain high-risk participants. For instance, given the increased use of smartphones and agents embedded in mobile phones by these vulnerable population, additional monitoring or assistive devices can be appropriated to aid any shortcomings of current technical, clinical, and social infrastructure that TIPS provides. Being able to increase retention among those who need the resources the most will be critical in maximizing the benefits TIPS can provide to its participants.

The role of caregivers as facilitators of home-based telehealth is critical. In community-based settings where the deployed telehealth system is multi-user and may be operated by support staff or a self-operated mobile unit (e.g., kiosk) we found greater retention when there was no caregiver, suggesting CTPs may support this critical role. Interestingly, our finding that participants without a caregiver reported greater odds of retention was limited to EP participants. LEP participants had inverse odds, though not significant, suggesting, they were not seeking a caregiver proxy. Rather, the significantly higher odds of LEP females remaining active in the program suggests, they are the caregivers. Accordingly, TIPS may serve as a great supportive mechanism for older adults who do not have caregivers, which consist the majority of Americans who are older adults. Furthermore, TIPS can be a supportive mechanism for LEP older adults who may have played multiple social roles in the household as caregiver, head of household, or women who have not had much support. TIPS recruited more LEP participants than most programs, but retained fewer LEP participants, demonstrating recruitment success but retention barriers. Previous studies suggest existing barriers to healthcare utilization among older LEP adults include health literacy and technology access (they don’t have the smartphones). These barriers also apply to CTPs, LEP participants struggle to engage in technology-based healthcare to the same extent of their non-LEP peers. There is a lack of evidence on the efficacy and feasibility of CTP participants with LEP. Studies on the acceptance of telehealth systems show usability as one of the main barriers to successful adoption. Specific to LEP populations are the lack of comprehensive usability and seamless integration into care delivery since language access is needed at every interaction, thus making accessibility a higher priority. Some studies show navigators help mitigate some of these barriers similar to the role that TIPS on-site technical support. On top of existing usability issues, the needs of LEP clients must be addressed to ensure their participation.

Pew Internet Research has shown older adults are increasingly using mobile phones and smartphones. Furthermore, due to the COVID pandemic, older adults are increasingly adopting mobile and computing devices. This rapidly changing technological adoption pattern encourages potential technological solutions that may ameliorate language barriers. Emerging studies show older adults’ increased use of conversational or voice-based agents, which have become prevalent in everyday mobile devices, due to their ability to generate more natural user interaction than existing user interfaces. These more accessible mobile applications specifically designed for low health and technological literacy and technological support for the TIPS staff to aid translation may generate solutions for the programs that lack resources around language interpretation and translation. Limitations of our study exist, starting with the larger proportion of female (N=1046) versus male participants (N=279). This gender difference, however, is not uncommon—the literature shows that older women are more likely than men to use technology and services. Though there were no significant differences by gender across language or attrition, future work should strive for a balanced sample. Further, MCCs and hospitalization history were self-reported and may not be fully accurate. But there is sufficient evidence of good concordance between self-reported data and medical record review. Finally, retention is measured as a binary outcome, more detail such as time in program could improve understanding of individual behaviors.

Conclusion
The rapid deployment of telemedicine interventions and the changing social and technological environment due to the COVID-19 pandemic brings urgency to fulfilling the needs of all populations without leaving behind the populations that can most benefit from technological innovations. Leveraging technology and community engagement can contribute to community resilience, especially in the face of a public health emergency. By supporting social interactions, community connectivity, and strong system resilience, the probability of successful aging-in-place is more likely. Our study occurred in the areas hardest hit by the initial wave of COVID-19 cases. Our study brings demonstrated viable solution to engaging older adults, while discovering critical challenges that the informatics community can address to support all populations.
References
Cross-Vendor CT Image Data Harmonization Using CVH-CT

Md Selim1,3, Jie Zhang, PhD2, Baowei Fei, PhD5,6, Guo-Qiang Zhang, PhD7, Gary Yeeming Ge2, Jin Chen, PhD1,3,4

1Department of Computer Science 2Department of Radiology 3Institute for Biomedical Informatics 4Department of Internal Medicine, University of Kentucky, Lexington, KY 5Department of Bioengineering, University of Texas at Dallas, Richardson, TX 6Department of Radiology, UT Southwestern Medical Center, Dallas, TX 7Department of Neurology, University of Texas Health Science Center at Houston, Houston, TX

Abstract While remarkable advances have been made in Computed Tomography (CT), most of the existing efforts focus on imaging enhancement while reducing radiation dose. How to harmonize CT image data captured using different scanners is vital in cross-center large-scale radiomics studies but remains the boundary to explore. Furthermore, the lack of paired training image problem makes it computationally challenging to adopt existing deep learning models. We propose a novel deep learning approach called CVH-CT for harmonizing CT images captured using scanners from different vendors. The generator of CVH-CT uses a self-attention mechanism to learn the scanner-related information. We also propose a VGG feature based domain loss to effectively extract texture properties from unpaired image data to learn the scanner based texture distributions. The experimental results show that CVH-CT is clearly better than the baselines because of the use of the proposed domain loss, and CVH-CT can effectively reduce the scanner-related variability in terms of radiomic features.

1 Introduction

Computed Tomography (CT) is one of the most commonly used imaging modalities for patient diagnostics due to its ability to capture detailed anatomical features1. For years, Canon (former Toshiba), Siemens, General Electric (GE), and Philips have been supplying high-quality CT scanners worldwide. Owing to increasing competition frequent innovations regarding slice count, dose optimization, reconstruction methods, etc. are taking place in the market. Each medical imaging vendor gradually develops its unique techniques2 to advance CT imaging. The divergence among CT imaging techniques, however, introduces image feature variations in terms of radiomic patterns2,3, which greatly hinder the progress of cross-vendor data sharing, ambiguity in large-scale data analysis, and automated diagnostics4.

To address the radiomic feature discrepancy problem, one of the computational approaches is to normalize the radiomic features of CT images captured with different protocols. However, since the concepts of radiomic features are not well defined5, an image feature normalization tool that works for a set of radiomic features may not work for the others. Alternatively, we can harmonize CT image data directly while preserving their anatomic details3,6. From the harmonized images, radiomics analysts can extract their desired radiomic features for further analysis without worrying about the feature discrepancy problem. Mathematically, let x be a CT image acquired using a scanner, \( \hat{x} \) be its corresponding image captured with a different scanner, the image harmonization framework aims to compose a synthetic image \( x' \) from \( x \), such that \( x' \) follows the feature distributions of \( \hat{x} \) rather than \( x \).

The CT image harmonization problem can be viewed either within scanner or between scanners. Recent progress on this topic has been focused on the former one. Choe et al7 developed a Convolutional Neural Network (CNN)-based approach for CT image standardization. The model learns the residual representation of the target images, and then a residual image is combined with its source image to generate a synthesized image. The model, since it trains a CNN from scratch, requires large training data. Liang et al8 proposed a cGAN-based9 CT image standardization model named GANai. An alternative training strategy was developed to effectively learn the data distribution. GANai achieved better performance comparing with cGAN and the traditional histogram matching approach10. However, GANai focuses on the relatively easier image patch synthesis problem rather than the whole DICOM image synthesis problem. Selim et al6 proposed a GAN-based CT image standardization model named STAN-CT. In STAN-CT, a loss function was developed to consider both the latent space loss and the feature space loss. While the former is adopted for the generator to establish a one-to-one mapping from standard images to synthesized images, the latter allows the discriminator to critic the texture features of both standard and synthesized images. Similar to GANai, STAN-CT was applied at image patches and only a few texture features were used as the evaluation criteria. Another GAN-based CT
The generators $A_2B$ and $B_2A$ are both CNN networks with a similar structure. The discriminator $D$ is a fully convolutional network classifier for determining the source of the synthesized images. A VGG texture feature-based domain loss is used to train the model with unpaired data. Domain loss is calculated as a mean square error between the gram matrix representation of $I_A$ and $I_B$.

Image standardization model named RadiomicGAN\textsuperscript{11} took the advantage of transfer learning and proposed a dynamic window-based training approach to adopt the learned information from the RGB image domain into the CT image domain. The results were evaluated on a wide range of radiomics features.

However, all these models require paired training data, greatly limiting the application scope to harmonizing images captured with the same type of scanners.

The cross-vendor CT image harmonization remains a critical bottleneck for inter-institutional data harmonization. This is mainly because it is difficult to obtain paired imagery data\textsuperscript{12}. For example, a patient is scanned using a GE scanner, it is less likely that the patient will be scanned using a Siemens scanner in close time. In this paper, we present a novel deep learning model, called CVH-CT (cross-vendor harmonization of CT images), for cross-vendor CT image harmonization (see Figure 1). CVH-CT relaxes the need for paired training data. Using unpaired training data, CVH-CT can synthesize images from vendor A to vendor B and vice versa. CVH-CT integrates a self-attention layer named CBAM (Convolutional Block Attention Map)\textsuperscript{13} to systematically learn the global features that appear in the images due to the use of different vendors. CVH-CT borrows the CycleGAN\textsuperscript{14} loss and improves it with the feature-based domain loss to determine the feature gap between the synthesized images and the target images in the target domain. Overall, CVH-CT has the following advantages:

1. CVH-CT effectively learns the feature distributions between two different CT image domains without paired training data.
2. The self-attention block, CBAM, with the convolutional layers helps the network to capture the scanner’s relevant global distribution.
3. The domain loss can calculate the feature gap between unpaired data which assists the network to learn the target domain distribution.
4. Experimental results show CVH-CT is better than the state-of-the-art CT image harmonization methods.

## 2 Background

**CT Scanner.** CT is a widely used clinical imaging modality for patient diagnostics. CT scanner vendors in the market, including Canon, Siemens, GE, and Philips, have developed unique reconstruction algorithms to improve image quality while reducing the dosage\textsuperscript{2}. The CT image radiomic features discrepancy problem due to the use of scanners with different imaging techniques poses a gap between CT imaging and radiomics studies\textsuperscript{3,3}. 
Radiomic Feature. Image features, commonly known as the radiomic features, are critical for radiomics study, e.g. tumor characterization\textsuperscript{15}. Mathematical and statistical models are used to extract these features from images. Radiomic features reflect the cellular and genetic levels phenotypic patterns that are hidden from the naked eye\textsuperscript{16,15}. Thus, there is a great potential to capture tumor heterogeneity and phenotypic details with radiomic features. However, the effectiveness of radiomic features, especially for large-scale cross-institute studies, is greatly reduced due to the non-standard practice of medical image acquisition\textsuperscript{17}, since radiomic features are dependent on both inter and intra-scanner protocol settings\textsuperscript{18,17}.

Image synthesis. Image synthesis is a common practice for artificial data generation which mimics the real data distribution\textsuperscript{19}. An image synthesis algorithm may require no input or a random noise for data synthesis but this type of algorithm has less control on the synthesized data\textsuperscript{20}. Medical data synthesis requires high quality with valid clinical meaning. So not all the available image synthesis models may not be appropriate for the medical domain. Recent progress on deep learning for image-to-image transformation provides better control on the synthesized data which makes a suitable choice to adopt it in the medical domain\textsuperscript{21}.

Generative Adversarial Networks (GAN), which are often used for data and image synthesis\textsuperscript{19}, normally consist of a generator $G$ and a discriminator $D$. The generator that could be a Convolutional Neural Network (CNN) is responsible for generating fake data from noise, and the discriminator tries to identify whether its input is drawn from the real or fake data. Among all the GAN models, cGAN is capable of synthesizing new images based on a prior distribution\textsuperscript{22}. The conventional GAN-based image-to-image translation model training requires paired training data that limits its scope where paired data is unavailable. A variation of GAN named CycleGAN\textsuperscript{14} is designed by combining two GAN models to synthesized images for unpaired data.

3 Methods

CVH-CT is designed to harmonize the scanner-related variability by generating synthesized CT images without the need for paired cross-vendor training data. Built on top of the CycleGAN framework, CVH-CT incorporates consecutive convolutional and self-attention layers to learn the global features of entire CT images efficiently. CVH-CT also introduces a new VGG\textsuperscript{23} feature-based domain loss function, which can be directly calculated using unpaired images.

CVH-CT Architecture

Based on the cycle generative adversarial network method called CycleGAN, CVH-CT consists of two generative adversarial networks (GANs). Each GAN has its generator $G$ and discriminator $D$ but with the same structure. A GAN tries to map the data distribution from data domain $A$ to domain $B$. The two generators are $G_{A\rightarrow B}: A \rightarrow B$ and $G_{B\rightarrow A}: B \rightarrow A$. The discriminator $D_A$ is responsible to distinguish between the data being real or fake in terms of domain $A$ and the same job is done by $D_B$ for domain $B$. In Figure 1, given training datasets from two domains, e.g. one set of CT images captured using scanner A and another set captured using scanner B, an initial mapping of $G_{A\rightarrow B}$ is learned to generate synthesized of domain B based on the input image from A. The mapping of $G_{B\rightarrow A}$ is learned to generate synthesized domain A’s image using the input image from B. Discriminator $D_A$ tries to determine whether the input image is from domain $A$ or not and the same for discriminator $D_B$.

CVH-CT Generator. The generator of CVH-CT, as shown in Figure 2, is a convolutional neural network with dedicated convolutional-attention layers. Specifically, each convolutional layer followed by an activation function is used to extract features from the previous layer and store it in a feature map. The convolutional layer dimension is kept consistent throughout the network. For every two convolutional layers (except the last three layers), a self-attention mapping is constructed to capture key global features from the corresponding convolutional layers. These self-attention layers are designed to learn the global features from the spatial dimension and the inter-channel dimension.

Convolutional Block Attention Map (CBAM). In complementary to the convolutional filters focused on extracting local features, self-attention layers are added into CVH-CT to capture the common features in each channel of the network. Such network structure change allows CVH-CT to capture domain-relevant features in the cross-vendor CT image harmonization problem. Figure 2 illustrates the architecture of CVH-CT where each self-attention layer called Convolutional Block Attention Map (CBAM)\textsuperscript{13} consists of the channel-wise attention and the spatial attention.

1101
Figure 2: The architecture of CVH-CT generator and discriminator. The generator uses a series of convolutional layers followed by a Convolutional Block Attention Map (CBAM). The convolutional layers are responsible for fetching the local details while the CBAM is responsible for fetching the global details within the feature map coming from its previous convolutional layers.

Channel-wise Attention Map: The channel-wise attention map is the first phase of CBAM that establishes an inter-channel relationship within a feature map. Max-pooling and avg-pooling operations are used to extract features from the spatial dimensions. Both feature maps are passed through a Multilayer Perceptron neural network (MLP). The output features are concatenated, and an element-wise product operation is used to produce the channel attention map.

Spatial Attention Map: The spatial attention map works on top of the output from the channel-wise attention map. Avg-pooling and max-pooling are used along with the channel axis. The average and maximum feature values pool out the impotent feature. Convolution is done on the concatenated feature of the average and max-pooling feature. After softmax, the output is element-wise multiplied with the input feature map. The output feature map is passed to the next layer for further process.

CVH-CT Loss

Conventional CycleGAN is trained using a comprehensive loss function including adversarial loss, cycle consistency loss, and identity loss. In CVH-CT, we introduce a new loss function called “domain loss” to compare key texture features between unpaired synthesized and target domain images.

Domain Loss. It is a widely used practice to extract style and content features using a pre-trained VGG network. In the cross-vendor CT image harmonization task, a synthesized image and an unpaired target image may have different anatomic structures but similar scanner-relevant texture features (aka. style). We adopted the style-related features from the trained VGG network, which are traditionally the layers closer to the fully connected layers in VGG. The domain loss is calculated as a Mean Square Error (MSE) of the gram-matrix representation of those layers. The domain loss is defined as:

$$ L_{fA} = \frac{1}{m} \sum_{i=1}^{m} (GRAM(G_{A2B}(a_i)) - GRAM(b_i))^2 $$  

$$ L_{fB} = \frac{1}{m} \sum_{i=1}^{m} (GRAM(G_{B2A}(b_i)) - GRAM(a_i))^2 $$

$$ L_f = L_{fA} + L_{fB} $$

where $GRAM(\cdot)$ is the gram-matrix calculated by multiply the feature maps by itself as described in Gatys et al. A and B are two image domains, and m is the batch size.
Adversarial loss. We apply the adversarial loss to both generators, where each generator synthesizes images of its target domain, while its corresponding discriminator tries to distinguish its input images from the target domain distribution or not. In model training, the generator aims to minimize the adversarial loss while its corresponding discriminator tries to maximize it. Note the discriminators if CVH-CT are trained only with $L$ distribution or not. In model training, the generator aims to minimize the adversarial loss while its corresponding discriminator tries to distinguish its input images are from the target domain Adversarial loss.

We apply the adversarial loss to both generators, where each generator synthesizes images of its target domain, while its corresponding discriminator tries to distinguish its input images are from the target domain.

$$L_{adv}(G_{A2B}, D_{B}, A) = \frac{1}{m} \sum_{i=1}^{m} (1 - D_{B}(G_{A2B}(a_i)))^2$$  \hspace{1cm} (4)

$$L_{adv}(G_{B2A}, D_{A}, B) = \frac{1}{m} \sum_{i=1}^{m} (1 - D_{A}(G_{B2A}(b_i)))^2$$  \hspace{1cm} (5)

where $m$ is the batch size, $a \in A$ is the input image of domain $A$ and $b \in B$ is the input image from domain $B$.

Cycle Consistency Loss. We apply the adversarial loss to both generators, where each generator synthesizes images of its target domain, while its corresponding discriminator tries to distinguish its input images are from the target domain distribution or not. In model training, the generator aims to minimize the adversarial loss while its corresponding discriminator tries to maximize it. Note the discriminators if CVH-CT is trained only with $L$.

$$L_{cycle}(G_{A2B}, G_{B2A}, A, B) = \frac{1}{m} \sum_{i=1}^{m} ||G_{B2A}(G_{A2B}(a_i)) - a_i|| + ||G_{A2B}(G_{B2A}(b_i)) - b_i||$$  \hspace{1cm} (6)

Identity Loss. The identity loss is introduced to preserve the key details of the input and target domains. It is calculated by proving the target domain data as input of the generator and calculating the L1 loss between input and synthesized images.

$$L_{Idt}(G_{A2B}, G_{B2A}, A, B) = \frac{1}{m} \sum_{i=1}^{m} ||G_{A2B}(b_i) - b_i|| + ||G_{B2A}(a_i) - a_i||$$  \hspace{1cm} (7)

Finally, the total loss of the CVH-CT generator $L(G)$ is defined as:

$$L(G) = L_{adv}(D) + \lambda_1 L_{cycle} + \lambda_2 L_{Idt} + \lambda_3 L_f$$  \hspace{1cm} (8)

where $\lambda_1 \in [0, 1]$, $\lambda_2 \in [0, 1]$ and $\lambda_3 \in [0, 1]$ are weight factors.

4 Experimental Results

Dataset

A multipurpose lungman chest phantom was scanned using the Siemens CT Somatom Force scanner and General Electric (GE) Revolution EVO scanner machine to obtain the training and testing data. The phantom has synthetic nodules in its lung which makes it an accurate life-size anatomical model of a human torso. Note that the phantom does not contain lung parenchyma, so the space within the vascular structure is filled with air. Using the human chest phantom with several synthetic lung nodules, in total 11,070 image slices were obtained from a GE scanner. In total 9,156 image slices were obtained using a Siemens scanner where 7,290 image slices are from the chest phantom and 1,866 are from lung cancer patients. All the image slices are 512 × 512 with 16-bit pixel encoding. The image acquisition parameters and the data description are in Table 1. Images in the training dataset are randomly paired. To evaluate the performance of CVH-CT and baseline models, we prepared a slice-by-slice paired dataset using phantom scans. Using the GE and Siemens scanners, in total 250 image pairs were generated with five different KVPs and a fixed (5 mm) slice thickness.

Also, we tested the performance of CVH-CT on CT images collected from the same scanner but with two different image reconstruction kernels using patient scans. The model was trained with a total of 9,580 CT image slices from lung cancer patients using four different slice thicknesses (0.5, 1, 1.5, 3mm) using the Siemens CT Somatom Force scanner. To enable model performance evaluation, we reconstructed two images from the same scan using reconstruction kernels Bl64 and Br40 respectively. Bl64 is the reconstruction kernel often used for lung screening, and Br40 is used for regular screening.
Table 1: Description of CT images acquired from GE and Siemens scanners.

<table>
<thead>
<tr>
<th>CT scanner</th>
<th>GE Revolution EVO</th>
<th>Siemens CT Somatom Force</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reconstruction Kernel</td>
<td>Lung</td>
<td>Bl64</td>
</tr>
<tr>
<td>Slice Thickness (mm)</td>
<td>0.625, 1.25, 2.5, 3.75 and 5</td>
<td>1, 1.5, 3 and 5</td>
</tr>
<tr>
<td>KVP</td>
<td>70, 80, 100, 120 and 140</td>
<td>70, 80, 100, 120 and 140</td>
</tr>
<tr>
<td>Total No. of Slices</td>
<td>11,070 (phantom)</td>
<td>7,290 (phantom), 1,866 (human)</td>
</tr>
</tbody>
</table>

Model Implementation

The generator of CVH-CT consists of seven convolutional layers and three CBAM blocks. Each convolutional layers have 64 filters with kernel size $4 \times 4$ and stride=1. The discriminator is a fully connected convolutional neural network with seven hidden layer. The convolutional layers used $4 \times 4$ filters. LeakyRelu is adopted as the activation function in all the hidden layers. The last layer of the generators uses $Tanh$ activation and discriminators use $Sigmoid$ activation. The model uses $80 \times 80$ soft-tissue image patches from the training dataset to train the model and the testing is done using the full CT slice that is $512 \times 512$. The synthesisization is done with the Hounsfield Unit (HU) ranging from $-1000$ to $900$. Random weights are used during the network initialization phase. Maximum training epochs are set to 50 with the learning rate being 0.0001 with momentum 0.5. The batch size is set to 32.

CVH-CT is implemented in PyTorch and runs on a Linux computer server with eight Nvidia GTX 1080 GPU cards. It takes about 4 hours to train the model from scratch. Once the model is trained, it takes about two seconds to harmonize a CT image slice. (Source code: https://github.com/A VAILABLE-SOON)

Evaluation Metric

Model performance was evaluated systematically at the whole image level and with randomly selected regions of interest (ROIs) from the soft tissues. For each CT image or ROI, quantitative radiomic features are extracted from six different feature classes using Pyradiomics. These feature classes and the number of features are, First Order Statistics (18), Gray Level Co-occurrence Matrix (GLCM, 24), Gray Level Run Length Matrix (GLRLM, 16), Gray Level Size Zone Matrix (GLSZM, 16), Neigbouring Gray Tone Difference Matrix (NGTDM, 5) and Gray Level Dependence Matrix (GLDM, 14). The six feature classes contain a total of 93 features.

We evaluated the CT image harmonization models by the radiomic features reproducibility analysis and the visual quality of the synthesized images. The radiomic feature reproducibility is analyzed by Concordance Correlation Coefficient (CCC) score and the visual quality of the synthesized images are determined by Peak Signal-to-Noise Ratio (PSNR), Structural SIMilarity (SSIM), and Normalized Cross Correlation (NCC).

Concordance Correlation Coefficient (CCC, see Eq. 9) was employed to measure the level of reproducibility of radiomic features. Mathematically, CCC represents the correlation between the input and the target image features in the six feature classes:

$$CCC = \frac{2\rho_{s,t}\sigma_s\sigma_t}{\sigma_s^2\sigma_t^2 + (\mu_s - \mu_t)^2}$$

where $\mu_s$ and $\sigma_s$ (or $\mu_t$ and $\sigma_t$) are the mean and standard deviation of the radiomic features belong to the same feature class in a synthesized (or target) image respectively, and $\rho_{s,t}$ is the Pearson correlation coefficient between $s$ and $t$. CCC ranges from -1 to 1 and is the higher the better.

Peak Signal-to-Noise Ratio (PSNR, see Eq. 10), which is a widely used metric for measuring the relative perceptual quality in image and video comparison. PSNR is defined as a logged ratio of the peak signal and the mean-square-error between the synthesized and the target images and is the higher the better.

$$PSNR = 10 \cdot \log_{10} \frac{MAX(X)^2}{\frac{1}{mn} \sum_{i=0}^{n} \sum_{j=0}^{m} (x_{ij} - x'_{ij})^2}$$

where $X$ and $X'$ are target and synthesized images with $m \times n$ dimension respectively. All the feature values are normalized to $[0, 1]$ range.
Table 2: CT image harmonization model performance comparison for images acquired with GE and Siemens scanners. The values represent the averaged (±standard deviation) CCC score for radiomic features of the synthesized images generated using different models.

<table>
<thead>
<tr>
<th>Feature Class</th>
<th>First order</th>
<th>GLCM</th>
<th>GLDM</th>
<th>GLRLM</th>
<th>GLSZM</th>
<th>NGTDM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input</td>
<td>0.89 ± 0.07</td>
<td>0.18 ± 0.15</td>
<td>0.28 ± 0.12</td>
<td>0.62 ± 0.16</td>
<td>0.31 ± 0.19</td>
<td>0.22 ± 0.2</td>
</tr>
<tr>
<td>CycleGAN</td>
<td>0.98 ± 0.02</td>
<td>0.24 ± 0.24</td>
<td>0.53 ± 0.44</td>
<td>0.74 ± 0.06</td>
<td>0.24 ± 0.21</td>
<td>0.41 ± 0.35</td>
</tr>
<tr>
<td>CVH-CT¹</td>
<td>1.00 ± 0.00</td>
<td>0.85 ± 0.14</td>
<td>0.89 ± 0.28</td>
<td>0.79 ± 0.21</td>
<td>0.41 ± 0.05</td>
<td>0.86 ± 0.18</td>
</tr>
<tr>
<td>CVH-CT²</td>
<td>1.00 ± 0.00</td>
<td>0.81 ± 0.23</td>
<td>0.85 ± 0.15</td>
<td>0.80 ± 0.18</td>
<td>0.77 ± 0.12</td>
<td>0.82 ± 0.13</td>
</tr>
<tr>
<td>CVH-CT³</td>
<td>1.00 ± 0.00</td>
<td>0.88 ± 0.11</td>
<td>0.88 ± 0.12</td>
<td>0.90 ± 0.14</td>
<td>0.72 ± 0.22</td>
<td>0.84 ± 0.16</td>
</tr>
</tbody>
</table>

Table 3: CT image harmonization model performance comparison for images acquired with GE and Siemens scanners. The values represent the averaged (±standard deviation) CCC score for radiomic features of the synthesized images generated using different models.

<table>
<thead>
<tr>
<th>Feature Class</th>
<th>First order</th>
<th>GLCM</th>
<th>GLDM</th>
<th>GLRLM</th>
<th>GLSZM</th>
<th>NGTDM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Siemens to GE conversion</td>
<td>0.89 ± 0.07</td>
<td>0.18 ± 0.15</td>
<td>0.28 ± 0.12</td>
<td>0.62 ± 0.16</td>
<td>0.31 ± 0.19</td>
<td>0.22 ± 0.2</td>
</tr>
<tr>
<td>CycleGAN</td>
<td>0.99 ± 0.01</td>
<td>0.60 ± 0.15</td>
<td>0.66 ± 0.18</td>
<td>0.47 ± 0.12</td>
<td>0.19 ± 0.18</td>
<td>0.73 ± 0.26</td>
</tr>
<tr>
<td>CVH-CT¹</td>
<td>1.00 ± 0.00</td>
<td>0.82 ± 0.22</td>
<td>0.81 ± 0.10</td>
<td>0.83 ± 0.15</td>
<td>0.79 ± 0.15</td>
<td>0.95 ± 0.15</td>
</tr>
<tr>
<td>CVH-CT²</td>
<td>1.00 ± 0.00</td>
<td>0.86 ± 0.12</td>
<td>0.80 ± 0.11</td>
<td>0.85 ± 0.10</td>
<td>0.80 ± 0.12</td>
<td>0.88 ± 0.11</td>
</tr>
<tr>
<td>CVH-CT³</td>
<td>1.00 ± 0.00</td>
<td>0.93 ± 0.28</td>
<td>0.87 ± 0.15</td>
<td>0.86 ± 0.00</td>
<td>0.75 ± 0.12</td>
<td>0.92 ± 0.18</td>
</tr>
</tbody>
</table>

Structural SIMilarity (SSIM, see Eq. 11) is defined as a co-relation between the synthesized and the standard images with values ranges from -1 to 1, and the value 1 indicates perfect structural similarity.

\[
SSIM = \frac{(2\mu_s\mu_t + c_1)(2\sigma_s\sigma_t + c_2)}{\mu_s^2 + \mu_t^2 + c_1(\sigma_s^2 + \sigma_t^2 + c_2)}
\]  

where \(\mu_s\) and \(\sigma_s\) (or \(\mu_t\) and \(\sigma_t\)) are the mean and standard deviation of the synthesized (or target) image respectively, and \(\sigma_{st}\) is the co-variance of \(s\) and \(t\). \(c_1\) and \(c_2\) are two constant. The SSIM values ranges between -1 to 1, and the value 1 indicates prefect structural similarity.

Normalized Cross Correlation (NCC, see Eq. 12), which is the correlation between a synthesized image and its corresponding target image and is defined as:

\[
NCC = \frac{\sum_{i=0}^{m} \sum_{j=0}^{n} x_{ij}x'_{ij}}{\sqrt{\sum_{i=0}^{m} \sum_{j=0}^{n} x_{ij}^2} \sqrt{\sum_{i=0}^{m} \sum_{j=0}^{n} x'_{ij}^2}}
\]

where \(X\) and \(X'\) are the target and synthesized image features respectively. All the values are normalized to [0, 1].

Performance Evaluation

We compared CVH-CT with the conventional CycleGAN and two CVH-CT variations. CVH-CT¹ is constructed with CycleGAN with the proposed domain loss \(L_f\). CVH-CT² is constructed with CycleGAN with the proposed CBAM layers. We evaluated all the models in two settings, i.e. transform a GE image into a synthesized image with the Siemens domain, and transform a Siemens image into a synthesized image with the GE domain. Note the source image of one domain is the target image of another domain. Table 2 shows the effectiveness of CT image harmonization in soft-tissue ROIs. The CCC scores of synthesized and actual images are reported for six different
Table 4: CT image harmonization model performance comparison for images acquired with GE and Siemens scanners. The values indicate the visual image quality of the harmonized images measured using PSNR, SSIM, and NCC.

<table>
<thead>
<tr>
<th></th>
<th>GE to Siemens conversion</th>
<th>Siemens to GE conversion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metrics</td>
<td>PSNR</td>
<td>SSIM</td>
</tr>
<tr>
<td>Input</td>
<td>24.11 ± 0.62</td>
<td>0.99 ± 0.02</td>
</tr>
<tr>
<td>CycleGAN</td>
<td>25.88 ± 0.14</td>
<td>0.99 ± 0.24</td>
</tr>
<tr>
<td>CVH-CT</td>
<td>26.62 ± 0.28</td>
<td>0.93 ± 0.44</td>
</tr>
<tr>
<td>CVH-CT²</td>
<td>26.50 ± 0.25</td>
<td>0.95 ± 0.04</td>
</tr>
<tr>
<td>CVH-CT³</td>
<td>26.75 ± 0.32</td>
<td>0.99 ± 0.06</td>
</tr>
</tbody>
</table>

Figure 3: Case study of model performance comparison. Heat maps of the residual images obtained by comparing two target images with the corresponding synthesized images under both GE-to-Siemens and Siemens-to-GE conversions. The input column includes images need to be harmonized and the rest of the columns are for the results of the respective models. Overall, CVH-CT has the lowest residues. Values close to 0 (blue) in heat map indicate high similarity between target and synthesized images, while values close to 1 (red) indicate high dissimilarity.

Feature classes. The row “Input” indicates the CCC scores of the source image in the six feature classes, which serve as the baseline. The rest CCC scores indicate the performance of all the compared models regarding the improvement of feature reproducibility. In both conversions, CVH-CT achieves the best score in five feature classes. All the values in the column “First order” is close to 1.0 indicating that anatomic features are well preserved with all the models. The CCC scores of CVH-CT are significantly higher than that of CycleGAN on the texture feature classes. Table 4 shows the PSNR, SSIM, and NCC scores for image quality assessment. Overall, CVH-CT effectively improved the visual quality in the synthesized images according to the three metrics.

We further tested CVH-CT using unpaired image data collected from the same scanner using two reconstruction kernels, namely Bl64 and Br40. The CCC score for Br40-to-Bl64 conversion is 0.87 ± 0.12 and the score for Bl64-to-Br40 conversion is 0.86 ± 0.10, indicating that CVH-CT works well not only on cross-vendor image harmonization but also on cross-protocol image harmonization.

Figure 3 shows the results of two sample images obtained from CycleGAN, CVH-CT¹, CVH-CT², and CVH-CT.
The source images of GE-to-Siemens conversion are the target images for Siemens-to-GE conversion and vice versa. The residual images at the second, fourth, sixth, and eighth rows were calculated as a numerical difference between a synthesized image and its corresponding target image. Values close to 0 (blue) indicate high similarity between them, while values close to 1 (red) indicate high dissimilarity between them. Overall, CVH-CT performed clearly better than CycleGAN. The residual image of CVH-CT indicates that its synthesized images are very close to the corresponding target images. For example, in the first row, the outer side of the lung has a high mismatch in all the synthesized images except for the image from the proposed CVH-CT model.

Figure 4 shows the effectiveness of the proposed domain loss in CVH-CT. In this case study, we synthesized an image in the GE domain using a Siemens image with different anatomic features (i.e. unpaired images). The visualization of images in different training epochs indicates that CVH-CT was efficiently trained to learn the textures from the GE domain. Also, these synthesized images at different epochs were not disturbed by the target image structure indicating the proposed domain loss can learn texture details while ignoring the structural information. The domain loss values are visualized in the right-bottom corner of the figure.

5 Conclusions

Data discrepancy in CT images due to the use of scanners with different imaging techniques adds an extra burden to radiologists and also creates a gap in large-scale cross-center radiomic studies. To facilitate large-scale medical image studies and to address the long-existing cross-center CT image data integration problem, we propose CVH-CT, a novel tool for CT image data harmonization. In CVH-CT, both the CBAM layers and the domain loss are introduced for efficient model training. CVH-CT can harmonize images without the need for paired training data. The experimental results demonstrate that CVH-CT is significantly better than the existing tools on CT image harmonization.

Acknowledgements

This research is supported by NIH NCI (grant no. 1R21CA231911) and Kentucky Lung Cancer Research (grant no. KLCR-3048113817).

References


Comparing Deep Learning and Conventional Machine Learning Models for Predicting Mental Illness from History of Present Illness Notations

Ingroj Shrestha, Padmini Srinivasan, Ph.D.
University of Iowa, Iowa City, Iowa, United States

ABSTRACT
Mental illness, a serious problem across the globe, requires multi-pronged solutions including effective computational models to predict illness. Mental illness diagnosis is complicated by the pronounced sharing of symptoms and mutual pre-dispositions. Set in this context we offer a systematic comparison of seven deep learning and two conventional machine learning models for predicting mental illness from the history of present illness free-text descriptions in patient records. The models tested include a new architecture CB-MH which ranks best for F1 (0.62) while another attention model is best for F2 (0.71). We also explore model decisions using Integrated Gradients interpretability method which we use to identify key influential features. Overall, the majority of true positives have key features appearing in meaningful contexts. False negatives are most challenging with most key features appearing in unclear contexts. False positives are mostly true positives in actuality as supported by a small-scale clinician-based user judgement study.

INTRODUCTION
Mental illness is the leading cause of disability globally – one in five adults in the U.S. lives with it[28]. Even prior to the COVID-19 pandemic mental illness and suicidal ideation were increasing among adults[32]. The same report shows that the number of people seeking help due to moderate or severe anxiety and depression has shown an alarming increase. These trends emphasize the importance of multi-pronged strategies to subdue the upsurge in mental health challenges: prediction models able to assist in effective diagnosis of mental illness is a key component.

The national drive towards Electronic Health Records (EHR) has resulted in vast collections of multi modal, structured and semi-structured patient records that compile information related to various encounters and activities with patients. These are naturally seen as resources that may be used to identify patients having specific diseases or disorders including those related to mental health. While EHR are generally encoded with disease codes from systems such as ICD-9/10/CM etc., there are well recognized limitations when these codes are used for secondary purposes such as phenotype or cohort identification. For example, these codes are known to be incomplete and error-prone[16, 31]. This has generated active research in building models that rely more on the free-text portions of patient EHRs, as for instance to predict disease states and prognosis. Thus we see papers and reviews on machine learning methods, both conventional and more modern deep learning (DL) for building prediction models from free-text EHR data[30, 31].

Despite the excitement over the development of DL models in health care applications[32], the emphasis on mental illness diagnosis from free-text portions of the clinical record has been minimal. In contrast, there has been active research with traditional machine learning models[2]. A 2020 review[28] of DL in mental health identifies under five papers exploring the free-text portions of EHR. These were directed at different goals including symptom severity[22] and cohort definition[23]. The one paper directly about mental illness prediction from free-text is by Tran et al[23]. They explore CNN and RNNs using the history of present illness portions of clinical notes for predicting 11 mental health conditions. The limitation is that they use a small dataset of only 1000 notes. Thus, it is still unknown as to how well deep learning models will perform on the task of predicting mental health conditions using free-text portions of EHR and consequently it is also unknown as to which DL models are best suited for this problem.

The problem of predicting mental illness is particularly challenging as unlike typical chronic conditions laboratory tests to aid in diagnosis are not generally available[25]. Instead there is reliance on patient observation, self reports, responses to questionnaires etc. These impressions are generally summarized in the patient notes. Another challenge is that the shared underlying biology is pronounced with mental disease. For example, people with depression share symptoms with those having generalized anxiety disorder such as irritability and sleep problems[17, 18]. Moreover, multiple problems may co-exist with one pre-disposing another[13]. Any predictive model will have to make fine distinctions between classes in this multi-label, multi-class setting.

1109
Set in this context, our goal is to rigorously compare deep learning (DL) methods in their ability to predict mental illnesses from free-text clinical notes. We also include conventional machine learning algorithms in our comparison. We systematically compare nine models representing CNNs, LSTMs, different mechanisms to include attention components (exploring both self and soft attention mechanisms), transformer based models as well as conventional SVMs and Logistic Regression. We also propose a DL model Convolution BiLSTM Multi-Headed attention that we call CB-MH. This architecture includes a multi-head attention mechanism. We compare the models using a large collection of 150,085 psychiatry clinical notes spanning 10 years (history of present illness portions). We make comparisons using both F1 and F2 measures. We then go deeper into one best DL model to get a better understanding of its predictions. For this we utilize an interpretability approach using the Integrated Gradients methodology which allows us to identify key features, with weights designating their relative importance in model decisions. We use these key features to examine reasons for true positive decisions as well as reasons for errors (both false positives and false negatives). We close with a small scale judgement study involving two clinicians who provide further insights on the nature of a sample of erroneous model decisions.

Next under Materials and Methods we describe our dataset, the DL and traditional ML models we compare. Next is the Results section followed by a section analyzing decisions made by a select DL model for one disease. Discussion, Conclusions and future work sections follow.

**MATERIALS AND METHODS**

**Dataset**

We used a 10-year span (2008-2017) of de-identified patient visit data from the psychiatry outpatient clinic of a leading midwestern hospital. There are 203,466 clinical visit records corresponding to 15,986 distinct patients. Each record has a de-identified patient id, visit date, notes on history of present illness (average length 219 words), and diagnoses in the form of ICD9 codes. After removing duplicates and records without a mental health diagnosis, there are 150,085 records for 14,916 patients. Domain experts grouped low frequency ICD labels resulting in a final set of 8 codes/diagnoses for our prediction goals. Table 1 describes our data.

Besides performing experiments on the Full set of 150,085 records we also experimented on a (Sample) of 15,009 records (10% of Full) to see if results scale well. Unlike most prior work, we explore this question while considering prediction confidence level. The smaller set is obtained using stratified sampling to maintain the same distribution of diagnosis as in the Full set (Table 1) within a 1% difference. The data were split into training (65%), validation (15%), and testing (20%). All records of a patient were either in training, validation, or testing.

**Deep Learning Models Tested**

**BERT:** This is a bidirectional language model based on the transformer model and is trained on two tasks masked language modeling (MLM) and next sentence prediction (NSP). MLM helps achieve bidirectionality and NSP learns sentence relationships during training. BERT is trained on a large corpus — the English Wikipedia and BooksCorpus and fine tuned for several NLP downstream tasks. We used pretrained uncased version of BERT base model modifying the last layer for our prediction task and fine tuning it.

**LSTM:** This is the standard sequence-to-sequence model (an extension of RNN) involving an input gate, forget gate, output gate, and a cell state to handle short term memory. The gates control the flow of input information (i.e., what information to remember or forget) and help to better handle context (in contrast with RNN). Equations representing these gates are as in Lipton et al. At an abstract level, we represent the final learned representation of the text as $\overrightarrow{h_t} = LST\overrightarrow{M}(x)$, where $x$ is the input clinical note, $\overrightarrow{h_t} \in R^{bh}$. where the hyperparameter $h_{\alpha}$ represents the size of

<table>
<thead>
<tr>
<th>Diagnosis</th>
<th>All</th>
<th>Train</th>
<th>Val</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unipolar Depression</td>
<td>51%</td>
<td>51%</td>
<td>49%</td>
<td>51%</td>
</tr>
<tr>
<td>Other</td>
<td>45%</td>
<td>46%</td>
<td>43%</td>
<td>45%</td>
</tr>
<tr>
<td>Anxiety Disorders</td>
<td>23%</td>
<td>23%</td>
<td>20%</td>
<td>23%</td>
</tr>
<tr>
<td>Substance Use Disorders</td>
<td>19%</td>
<td>18%</td>
<td>23%</td>
<td>19%</td>
</tr>
<tr>
<td>PTSD_Osc_PanicDisorder</td>
<td>19%</td>
<td>19%</td>
<td>16%</td>
<td>21%</td>
</tr>
<tr>
<td>Cluster B Personality Traits</td>
<td>16%</td>
<td>16%</td>
<td>16%</td>
<td>17%</td>
</tr>
<tr>
<td>Psychotic Disorders</td>
<td>14%</td>
<td>14%</td>
<td>16%</td>
<td>12%</td>
</tr>
<tr>
<td>Bipolar Disorders</td>
<td>12%</td>
<td>11%</td>
<td>13%</td>
<td>14%</td>
</tr>
</tbody>
</table>

Table 1: Data distribution of Full set with 150,085 records (Since a record can have multiple class labels the columns do not add up to 100%).
hidden layer.

**BiLSTM:** Here two LSTMs are combined, one that processes the text in the forward direction and the other in the backward direction. Both the LSTMs are of the same dimensions. Representing text \( x \) processed in the reverse direction as: \( \hat{h}_t = \overrightarrow{\text{LSTM}}(x) \), the final representation \( (H) \) is as in Eq.1 where \( H \in R^{2h_n} \).

\[ H = [\overrightarrow{h}_t, \overleftarrow{h}_t] \quad (1) \]

Based on insights from He et al.\(^6\) and preliminary experimentation with the validation data, we found it effective to add a global max pooling layer followed by a dense layer with ReLU activation before the final output layer.

**CNN:** This is a standard CNN with convolutional filters and max pooling followed by a dense layer with a non-linear activation function ReLU\(^9\).

**CNN-BiLSTM:** The objective here is to take advantage of the CNN convolutional layer and BiLSTM hidden representations. Thus we built this model by adding a BiLSTM on top of the CNN.

**CB-Atn:** This is our first model exploring attention. Here we add to the top of the CNN-BiLSTM architecture a soft attention layer\(^1\). We use \( H \) from (Eq.1) to obtain word level attention weights \( (a) \) using Eq.2b. These activation weights indicate the relative importance of different portions of the input text. The final vector \( (f) \) representation weighted by attention uses Eq.2c where \( W \in R^{c_v \times 2h_n}, b \in R^{c_v} \), and \( v \in R^{c_v} \) is a learned context vector.

\[
\begin{align*}
  u &= \tanh (WH + b) \\
  a &= \frac{\exp (u^T v)}{\sum_T \exp (u^T v)} \\
  f &= \sum_T aH
\end{align*}
\]

**CB-MH:** Inspired by work with the transformer model\(^2^4\) we propose a new architecture with multi-head self-attention mechanism instead of the single head soft attention. A single head will get different distributions each time it is run. Our insight is that a multi-head architecture will reduce this variation in distributions, thus giving attention to similar phrases. Thus, we capture attention multiple \( (r) \) times with different weight matrices \( W^Q_i, W^K_i, W^V_i \) for \( i = \{1, 2, ..., r\} \) with a final concatenation of attention using Eq.3a. Like CB-Atn this model builds off CNN-BiLSTM.

\[
\begin{align*}
  MultiHead(Q, K, V) &= \text{Concat} (\text{head}_1, ..., \text{head}_r) W^0 \\
  \text{head}_i &= \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \\
  \text{Attention}(Q, K, V) &= \text{softmax} \left( \frac{QK^T}{\sqrt{d_k}} \right) V
\end{align*}
\]

Here \( W^0 \in R^{d_v \times 2^h_n}, W_i^Q, W_i^K, W_i^V \in R^{2^h_n \times d_k} \) are learned parameters, and \( d_k = d_v \), where \( d_v \) represents the hidden size per head. The configuration for hidden size and the number of attention heads is the same as in BERT\(_{\text{BASE}}^4\). In Eq.3c the attention function maps a set of queries \( (Q) \), key-value pairs \( (K-V) \) to an output. In our case, we set the values of \( Q, K, \) and \( V \) as the output of BiLSTM layer (Eq.1), i.e. \( \text{MultiHead} (H, H, H) \). Our code for CB-MH is available at [https://github.com/IngrojShrestha/CB-MH](https://github.com/IngrojShrestha/CB-MH).

All models were configured with optimal hyperparameter settings using validation data. Table2 lists the key parameters explored and the selected values.
Table 2: Parameters for deep learning models (tuning range for hidden layer: [64, 128, 256], embedding dim.: [200, 400], dropout: [0.5, 0.6, 0.7, 0.8], global output threshold: [0.2, 0.4, 0.5, 0.6, 0.8], kernel size: [3, 4, 5, (3, 4, 5), (2, 3, 4, 5)]. dropout is used in penultimate layer.

<table>
<thead>
<tr>
<th>Model</th>
<th>Parameters (Optimal values selected)</th>
</tr>
</thead>
<tbody>
<tr>
<td>General</td>
<td>embedding dim. (400), learning rate (0.001), batch size (100), IG (m = 10) (\text{[24]}), global output threshold (0.2), epoch (10)</td>
</tr>
<tr>
<td>BERT(_{\text{BASE}})</td>
<td>default from Devlin et al.(\text{[3]})</td>
</tr>
<tr>
<td>LSTM</td>
<td>hidden layer size (128)</td>
</tr>
<tr>
<td>BiLSTM</td>
<td>hidden layer size (128), patience (2), L2 regularization ((\lambda = 0.0001), dense(128)</td>
</tr>
<tr>
<td>CNN</td>
<td>kernel size (2,3,4,5), stride (1), pool size (4), L2 ((\lambda = 0.0001)), dropout(0.5), dense(128)</td>
</tr>
<tr>
<td>CNN-BiLSTM</td>
<td>CNN kernel size (3), dropout (0.7), dense (128)</td>
</tr>
<tr>
<td>CB-Atn</td>
<td>(c_v) (128), dropout (0.7), dense (128)</td>
</tr>
<tr>
<td>CB-MH</td>
<td>attention head (12), hidden size (768), dropout (0.7)</td>
</tr>
</tbody>
</table>

**General settings for DL models:** The input texts had stopwords removed and were stemmed. We use pre-padding and pre-truncation to handle variable sequence length. Input text tokens were represented by 400 dimension pre-trained embedding vectors using word2vec\(\text{[23]}\) trained on PubMed abstracts from PubMed Baseline 2018\(\text{[29]}\). The final output layer (of size \(L\) equal to the number of classes) uses a sigmoid function to output the probability of prediction for each label \(P(\cdot)\). All DL models learn the optimal parameters by minimizing the binary-cross entropy loss in Eq. 4. We use the Adam optimizer with a learning rate of 0.001 and Keras for implementation using one NVIDIA Tesla P100 PCIE (16GB) GPU.

\[
\left\{-\frac{1}{L} \sum_{i=1}^{L} y_i \log P(y_i = 1) + (1 - y_i) \log (1 - P(y_i = 1))\right\}
\]

(4)

**Traditional Machine Learning Models Tested**

We test (1) Logistic Regression and (2) SVM with L2 regularization. Each note is represented by a TF-IDF weighted word unigram vector after stopword removal. For LR no stemming is optimal while for the SVM stemming (Porter’s stemmer\(\text{[26]}\)) is optimal. We create one SVM classifier/LR regressor for each class using scikit-learn\(\text{[20]}\).

**Performance Measures**

We report micro-averaged recall (R), precision (P), F1 and F2 scores. F2 (recall twice as important as precision) is important in medicine to minimize risk of missing a positive, i.e., of making Type II errors. Micro-averaging is appropriate as each visit is considered equally important.

**MODEL COMPARISON RESULTS**

Table 3 presents the results. We use bootstrapping shift test\(\text{[25]}\) to estimate 95% confidence intervals for comparing models. Specifically, we sample the test set with replacement 10K times; for each sample we calculate both F1 and F2 per model and we then generate 95% confidence intervals over these 10K bootstraps for each model and measure.

We see that F1 and F2 scores are significantly higher in the *Full* dataset compared to in *Sample* with only four exceptions of 18 comparisons (e.g., exception: F2 scores for our model CB-MH are statistically equivalent in both datasets). 18 comparisons refers to 9 models * 2 scores (F1 and F2). For some cases on moving from *Sample* to *Full* (a) F1 improved but F2 worsens (b) F1 improvements are significant but F2 improvements are marginal. When we move from *Sample* to *Full*, precision (P) improved but recall (R) decreased for some models. Moreover, the magnitude of change in precision was relatively large compared to that for recall in the 2 cases (out of 9) where F2 drops while F1 increases. Most importantly, confidence intervals for *Full* are tighter than for *Sample* and the DL models better distinguishable. For instance, in *Sample*, CB-MH and BiLSTM do not have significantly different scores, whereas in *Full* they do. This is consistent with expectations as DL models are more stable given more data. Thus, in our remaining analyses we focus on the *Full* dataset results.
With one exception (LSTM in F1), all DL models are significantly better than SVM and LR in both F1 and F2. The smallest statistically significant improvements are 3.6% in F1 (CNN) and 12.2% in F2 (LSTM), the highest 12.7% in F1 (CB-MH) and 44.9% in F2 (CB-Atn).

We find that F2 scores (minimizing Type II errors) are systematically larger than F1 scores for the DL models. But for SVM and LR, F1 scores are better than F2, these appear to favor precision over recall. But as noted earlier these models do not do as well as the DL models.

The attention models perform the best. **CB-MH** is the best for F1 (0.62 (95% CI : [0.613, 0.618])). As described earlier this involves a multi-head attention (12 heads) architecture. For F2 the best is CB-Atn of 0.71 (95% CI : [0.705, 0.710]). This uses a soft attention component. While our CB-MH performance is close (0.68 F2), there is a small difference of 0.02 in their confidence intervals. The use of multi-head attention seems to provide an advantage in making more precise predictions with a penalty in recall which brings down F2. BiLSTM and CNN-BiLSTM are second best on both F1 and F2 and LSTM is the weakest. The unidirectional sequence based model (LSTM) is not better than the non-sequential CNN. On the other hand, bidirectionality helps make BiLSTM better than CNN. Also, BERT trained on Wikipedia and Book-Corpus is likely limited for biomedicine.

**Table 3**: Results with 95% confidence intervals (*p*-value < 0.05)

<table>
<thead>
<tr>
<th>Model</th>
<th>Sample set F1</th>
<th>F2</th>
<th>Full set F1</th>
<th>F2</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>0.50 [0.492, 0.516]</td>
<td>0.47 [0.454, 0.478]</td>
<td>0.54 [0.534, 0.541]</td>
<td>0.49 [0.488, 0.496]</td>
</tr>
<tr>
<td>LR</td>
<td>0.51 [0.494, 0.518]</td>
<td>0.44 [0.436, 0.457]</td>
<td>0.55 [0.547, 0.554]</td>
<td>0.49 [0.491, 0.498]</td>
</tr>
<tr>
<td>BERT</td>
<td>0.57 [0.559, 0.576]</td>
<td>0.68 [0.670, 0.688]</td>
<td>0.58 [0.576, 0.582]</td>
<td>0.645 [0.642, 0.648]</td>
</tr>
<tr>
<td>LSTM</td>
<td>0.53 [0.526, 0.543]</td>
<td>0.63 [0.624, 0.644]</td>
<td>0.54 [0.533, 0.539]</td>
<td>0.55 [0.548, 0.555]</td>
</tr>
<tr>
<td>BiLSTM</td>
<td>0.56 [0.553, 0.571]</td>
<td>0.68 [0.668, 0.687]</td>
<td>0.60 [0.597, 0.602]</td>
<td>0.70 [0.694, 0.700]</td>
</tr>
<tr>
<td>CNN</td>
<td>0.52 [0.513, 0.532]</td>
<td>0.57 [0.561, 0.582]</td>
<td>0.57 [0.570, 0.576]</td>
<td>0.65 [0.645, 0.651]</td>
</tr>
<tr>
<td>CNN-BiLSTM</td>
<td>0.54 [0.531, 0.548]</td>
<td>0.68 [0.671, 0.688]</td>
<td>0.60 [0.599, 0.604]</td>
<td>0.70 [0.696, 0.701]</td>
</tr>
<tr>
<td>CB-MH</td>
<td>0.56 [0.551, 0.567]</td>
<td>0.68 [0.671, 0.688]</td>
<td>0.62 [0.613, 0.618]</td>
<td>0.68 [0.678, 0.685]</td>
</tr>
<tr>
<td>CB-Atn</td>
<td>0.55 [0.543, 0.560]</td>
<td>0.67 [0.658, 0.675]</td>
<td>0.595 [0.593, 0.598]</td>
<td>0.71 [0.705, 0.710]</td>
</tr>
</tbody>
</table>

**Table 4**: Classwise macro performance (CB-MH)

<table>
<thead>
<tr>
<th>Diagnosis</th>
<th>P</th>
<th>R</th>
<th>F1</th>
<th>F2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unipolar Depression</td>
<td>0.58</td>
<td>0.96</td>
<td>0.72</td>
<td>0.85</td>
</tr>
<tr>
<td>Anxiety Disorders</td>
<td>0.42</td>
<td>0.56</td>
<td>0.48</td>
<td>0.53</td>
</tr>
<tr>
<td>PTSD_OCD_PanicDisorder</td>
<td>0.57</td>
<td>0.51</td>
<td>0.54</td>
<td>0.52</td>
</tr>
<tr>
<td>Substance Use Disorders</td>
<td>0.56</td>
<td>0.47</td>
<td>0.51</td>
<td>0.49</td>
</tr>
<tr>
<td>Cluster B Personality Traits</td>
<td>0.42</td>
<td>0.63</td>
<td>0.51</td>
<td>0.57</td>
</tr>
<tr>
<td>Psychotic Disorders</td>
<td>0.73</td>
<td>0.74</td>
<td>0.73</td>
<td>0.74</td>
</tr>
<tr>
<td>Bipolar Disorders</td>
<td>0.60</td>
<td>0.42</td>
<td>0.49</td>
<td>0.45</td>
</tr>
<tr>
<td><strong>macro averages (CB-MH)</strong></td>
<td><strong>0.55</strong></td>
<td><strong>0.61</strong></td>
<td><strong>0.57</strong></td>
<td><strong>0.59</strong></td>
</tr>
<tr>
<td><strong>macro averages (LR)</strong></td>
<td><strong>0.70</strong></td>
<td><strong>0.37</strong></td>
<td><strong>0.48</strong></td>
<td><strong>0.40</strong></td>
</tr>
</tbody>
</table>

Table 4 provides class level results (ignoring Other class) four **CB-MH** with rows ordered by decreasing data size. With one strong exception (Psychotic disorders), we see higher F1 with larger class size. The pattern is less consistent for F2 due to differences in the relative success with recall and precision. E.g., recall is favoured in Unipolar Depression and Cluster B personality and precision in Bipolar disorders. Exploring influential factors for these patterns in recall over precision (beyond dataset size variations) is left to future work. For completeness, we report macro-averages across classes and for the best traditional machine learning approach (LR). The percentage improvement of our proposed model over the LR approach is 18.8% in macro F1 and 47.5% in macro F2; these percentage improvements are in fact larger than micro average improvements reported earlier (12.7% F1, 38.8% F2, Table 3).
ANALYSIS OF CORRECT AND ERRONEOUS MODEL PREDICTIONS

We analyze model predictions using Integrated Gradients (IG) an interpretability or explainability technique. We analyze both correct decisions (focusing on true positives (TP)) as well as erroneous decisions, both false positive (FP) and false negative (FN) decisions (see Error Analysis section). We focus our analysis on decisions made by the CB-Atn model which had the highest F2 scores and the class unipolar depression. Our CB-MH model is an equally viable model for analysis since it leads in F1. However, we chose to analyze the model with lower Type II errors. First we present an overview of the IG based analysis approach and then present our observations.

Integrated Gradients (IG) estimates the extent to which each word feature contributes to model predictions. This is estimated based on the gradients of the model with respect to the input. At a high level the method operates as follows. We are given the embedding vector representation ($e$) for the input text, the baseline zero embedding vector ($e'$), and $F(.)$ the function in the model that gives the output vector. The interpolated method computes a series of $m$ vectors (here $m = 100$) between $e$ and $e'$. Then gradients for each interpolated vector are computed; the gradient for an input feature is the partial derivative of $F(.)$ with respect to that feature ($e_i$). The final integrated gradient score for a word feature ($IG_i$) is then the average of its $m$ gradients multiplied by $(e_i - e_i')$. This calculation is as in Sundararajan et al. and is shown in Eq. 5.

$$IG_i = (e_i - e_i') \sum_{k=1}^{m} \frac{\partial F(e' + \frac{k}{m} (e - e'))}{\partial e_i} \frac{1}{m}$$

IG scores are normalized $(-\infty, +\infty)$ to $[0, 1]$ with negative values mapped to $[0, 0.5]$ and positives to $(0.5, 1]$. A feature with a positive value indicates that its presence increases the likelihood of of the model predicting a label. Similarly, a negative value indicates that the feature’s presence in the note reduces the likelihood of the same event. Scores close to 0.5 signal neutral features.

Analysis strategy

Given a note of interest we first use IG to identify key features used by the CB-Atn model to make its predictions. Key feature(s) is/are the word(s) with the highest IG score. We limit features to relevant UMLS semantic types: Mental or Behavioral Dysfunction, Sign or Symptom, Disease or Syndrome, Mental Process, and Pharmacologic Substance. Our code for semantic filtering is available at [https://github.com/IngrojShrestha/CB-MH](https://github.com/IngrojShrestha/CB-MH). We analyze a note if it has a key feature with IG score $\geq 0.51$, i.e., it has some positive influence on the prediction. The two authors then manually categorized a surrounding context of $k$ ($=15$) words into: (1) positive: context indicates depression present, (2) negative: context indicates depression absent, or (3) unclear: context insufficient to analyze. There was strong agreement between the two sets of judgements (kappa=0.81; 89.9% agreement).

Analysis of True Positives

There were 113 TP notes to analyse. The majority (82%) appear in positive contexts indicating sound reasons for these correct positive decisions. We see such examples in Table 5. Notably, in the last example the patient has depression that is improving making this a correct prediction. While depression appears as a key feature in some notes it is crucial to note that a simple keyword search strategy using ‘depression’ on our Full dataset yields very poor results (F1: 0.47, F2: 0.41). The remaining 18% of TP key features appear in unclear/uninterpretable contexts. Thus the majority of the true positive decisions made by the model are sensible and justifiable.
Table 5: True positive examples (key feature is in bold)

<table>
<thead>
<tr>
<th>Key Feature</th>
<th>Context</th>
</tr>
</thead>
<tbody>
<tr>
<td>depressive</td>
<td>a 44 year old white male with a history of major depressive disorder extending back to an original episode in 2005 tried on several different medications including</td>
</tr>
<tr>
<td>exhausted</td>
<td>to take the next dose she has significant anxiety in addition to depression and feels exhausted she said that part of her sleep problems come from the fact that her chronic</td>
</tr>
<tr>
<td>disorder</td>
<td>mg by mouth daily for 2 weeks then 2 tablets daily thereafter indications major depressive disorder previous medications acetaminophen tylenol 500 mg tablet take 500 mg by mouth 4 times daily</td>
</tr>
<tr>
<td>depression</td>
<td>45 y o y o caucasian woman presents for evaluation and treatment of depression pt reports she is doing well pt reports good mood sleep appetite interest and concentration</td>
</tr>
</tbody>
</table>

Analysis of False Positive and False Negative Errors

False Positives: Key features occur in two types of contexts

The 34 FP notes analyzed were about evenly split between positive for depression contexts and unclear contexts. In the first example in Table 6, the patient is under antidepressants, which indicates that model prediction is correct. The next two examples indicate that depression continues from the past into the present. In the last example, while the patient is improving treatment for depression is clearly stated. The interpretations for these FP decisions (and we found several similar ones) point to likely errors of omission in ICD annotations.

Table 6: False positive examples (key feature is in bold)

<table>
<thead>
<tr>
<th>Key Feature</th>
<th>Context</th>
</tr>
</thead>
<tbody>
<tr>
<td>emotion</td>
<td>not want to continue antidepressants because she reported that she is less control over her emotions she recently was in the process of applying for a part time job for approximately</td>
</tr>
<tr>
<td>depression</td>
<td>name patient states that she was first diagnosed and put on medications for depression in 1991 she was hospitalized once in 1996 she states that typically depression manifests as</td>
</tr>
<tr>
<td>disorder</td>
<td>that encounter are available the patient reports that she has been diagnosed with major depressive disorder and has been treated for that for nearly 20 years she presents today requesting refills</td>
</tr>
<tr>
<td>insomnia</td>
<td>some overall improvement review of depressive symptoms shows that she has some initial and intermittent insomnia and fatigue particularly in the early morning she feels she should resume her previous morning</td>
</tr>
</tbody>
</table>

False negatives: Key features are mostly in unclear contexts

The vast majority (84%) of the 74 FNs analyzed fall into unclear contexts making this the most challenging class of model errors for analysis. In example 1 in Table 6, the patient note has: “...Trazodone causes severe daytime sedation so he avoids it... .... He agrees to reduce Remeron to 7.5 ...” The model fails to connect Remeron and Trazadone with unipolar depression, possibly due to insufficient training data. Instead, it views thoughts as the key feature but its context is not specific enough to predict unipolar depression; instead Bipolar Disorders is predicted. In the (psychosis) example the model clearly made an error. In example three, the context for anxiety indicates the absence of depression (i.e., negative context). This could be why the model does not predict unipolar depression. Examining the full note we do not find any indication of a patient having depression. Similarly, in the last example (paranoia) there is no evidence in the rest of the note. Thus in a few cases the IG interpretation shows that evidence for a label is missing, pointing to possible errors of commission in ICD labeling.

Error Analysis with Clinical Expertise

Next we present a small user study to determine if the model’s FNs and FPs decisions are due to annotation errors.

Sample selection: We first ranked the FP and FN notes by IG score given to a key feature. To explain, the top ranking
Table 7: False negative examples (key feature is in bold)

<table>
<thead>
<tr>
<th>Key Feature</th>
<th>Context</th>
</tr>
</thead>
<tbody>
<tr>
<td>thoughts</td>
<td>having episodes like this in the past which resolve on their own he denies racing thoughts or worries at night trazodone causes severe daytime sedation so he avoids it once asleep</td>
</tr>
<tr>
<td>psychosis</td>
<td>on her husband's part the differential diagnosis was considered major depression with psychosis versus psychosis nos and started on zoloft titrated rapidly to 100 mg and risperdal 1 mg qhs</td>
</tr>
<tr>
<td>anxiety</td>
<td>davies and has a key chain and jacket no evidence of depression psychosis mania or anxiety no alcohol or drugs past psychiatric history</td>
</tr>
<tr>
<td>paranoia</td>
<td>no current depressive sx does not believe he currently has an anxiety problem no hallucinations paranoia no history of mania past psychiatric history</td>
</tr>
</tbody>
</table>

Manual annotation: Two clinicians independently annotated each note. They were given the full history of present illness and asked to decide if an ICD code pertaining to depression should or should not be assigned. They were also allowed to use a ‘not sure’ category. They agreed in 18/25 cases (kappa = 0.59; 85.7% agreement). Four notes had at least one judge mark ‘not sure’ while they disagreed on three notes. Gauged against the 18 decisions where the judges agreed, our model’s FP and FN decisions were correct in 78% (14/18). To explain, we marked an FP (FN) decision as correctly declared positive (negative) if both judges indicated that the note should (should not) be annotated with a depression ICD code. Interestingly, all FP decisions were declared correct (12/12) while a third (2/6) FN decisions were correct. These results are encouraging especially for the FP decisions.

DISCUSSION

Our proposed model, \( CB-MH \), is top ranked in F1 (0.62) while \( CB-Atn \) is best in F2 (0.71). The differences between their confidence intervals is slight (0.02 at most). Both models use attention. In the case of our model the run time for both training and testing with the Full set is approximately 122 minutes. We find that while scores are generally higher in the Full set compared to the Sample, notably confidence intervals are tighter with the larger set. Traditional machine learning algorithms do not fair as well as DL models. They seem less well equipped to capitalize on the semantics underlying diagnoses as well as DL models. They seem less well equipped to capitalize on the semantics underlying diagnoses as well as DL models. DL models are good at reducing Type II errors achieving higher F2 scores compared to F1 scores. As pointed in a recent review of DL models operating off clinical texts, it is important to continue comparing DL with traditional methods. While our DL models fared well, in 11% of the studies they surveyed DL algorithms performed worse than traditional machine learning algorithms.

Setting aside differences across papers, our best scores are lower than recent DL results for heart failure (F1: 0.9), COPD (F1: 0.9), and kidney diseases (F1: 0.92). Our LSTM (F1:0.54) also fared lower than an LSTM for cardiovascular disease (F1:0.61). These differences underline the greater challenge in predicting mental illness with its fine distinctions between disease classes. Our results also point to room for improvement and the need for research on stronger decision models for this challenging problem setting.

A high majority (82%) of true positive model decisions for depression analyzed involve key features that appear in meaningful contexts. This points to model success in this regard. Interestingly, about half of the FPs analyzed involve features in contexts indicating disease presence - suggesting possible errors of omission in ICD labels. Our follow up small scale study using clinician judges lend further support to this suggestion. All of the model’s FPs reviewed by the judges using the notes where declared to be true positives. The FN group was the most challenging: 84% of analyzed notes had key features that were in unclear contexts. For some notes, the model was clearly incorrect, possibly there were weaknesses in the training data. In still others, evidence supporting a positive annotation for depression was not in the note. In this context, the clinician judges found a third of the model's FN decisions to be true negatives. Overall, agreement between judges was moderate (clinician judgements) to strong (author judgements). This suggests - in general - that the task of looking for evidence supporting depression diagnosis, either in a text window of +/-15 words or in the whole note is more straightforward than nuanced.
Our error analysis, especially with the false positives indicate a possibility for the future. Perhaps there is the potential for combining model prediction from free-text with interpretability based error analysis as a pathway towards strengthening, i.e., selectively correcting ICD annotations of clinical records. We will explore this in future research.

CONCLUSION

We compared seven deep learning and two traditional machine learning algorithms on the multi-class, multi-label problem of predicting mental illness from free-text notes. The tested models included one that we proposed CB-MH, a multi-head, soft-attention model. This model achieved the highest F1 score and a close to highest F2 score. Error analysis was conducted using Information Gradient methodology to identify features which contributed most to model decisions. We showed through clinician judgements that FPs with highest IG scoring features were actually TPs; similarly a third of the FNs were true negatives. We show a potential way forward for using DL models with error analysis through IG as a mechanism to correct ICD annotation errors in clinical datasets.

LIMITATIONS AND FUTURE WORK

We have not explored resources such as BioBERT. Though built for applications related to genes and diseases these may prove useful. Given the manual effort involved we focus on extracting explanations for a single disease. Our next goal is to scale this up by automating the process of identifying explanations from IG scores of features and their contexts. We analyzed micro averaging results alone (Table 3) primarily due to space restrictions. With micro averaging larger classes are likely to dominate. However, Table 4 does provide some results with macro averaging.

ACKNOWLEDGEMENTS

The authors thank the two clinicians Dr. L. Durairaj and Dr. A. Aggarwal for their participation as judges in the user study.

References

BlockIoT: Blockchain-based Health Data Integration using IoT Devices

Manan Shukla¹, Jianjing Lin, PhD¹, Oshani Seneviratne, PhD¹
¹Rensselaer Polytechnic Institute, Troy, NY, USA

Abstract

The development and adoption of Electronic Health Records (EHR) and health monitoring Internet of Things (IoT) Devices have enabled digitization of patient records and has also substantially transformed the healthcare delivery system in aspects such as remote patient monitoring, healthcare decision making, and medical research. However, data tends to be fragmented among health infrastructures, and prevents interoperability of medical data at the point of care. In order to address this gap, we introduce BlockIoT that uses blockchain technology to transfer previously inaccessible and centralized data from medical devices to EHR systems, which provides greater insight to providers who can, in turn, provide better outcomes for patients. This notion of interoperability of medical device data is possible through an Application Programming Interface (API), which serves as a versatile endpoint for all incoming medical device data, a distributed file system that ensures data resilience, and knowledge templates that analyze, identify, and represent medical device data to providers. Our participatory design survey on BlockIoT demonstrates that BlockIoT is a suitable system to supplement physicians’ clinical practice and increases efficiency in most healthcare specialties, including cardiology, pulmonology, endocrinology, and primary care.

Introduction

In the last several decades, the explosion of information technology has profoundly changed how information is stored, exchanged, managed, and analyzed in medicine. For instance, the invention of EHRs has enabled digitization of patient records and has substantially transformed the healthcare delivery system. At the same time, the emergence of IoT health devices has also been increasingly applied in remote patient monitoring, healthcare decision making, and medical research by offering a variety of physiological data (such as heart rate, EKG, blood sugar levels) of the patient [1]. The advancement of these technologies has created the potential to improve the quality and efficiency in the healthcare industry. However, clinical data is still primarily stored and managed in a fragmented manner, which creates friction in information exchange at the point of care and also hinders large-scale health-data research empowered by technology such as artificial intelligence/machine learning (AI/ML). This paper focuses on the integration between EHRs and health-monitoring IoT devices in a decentralized environment, in the hope of offering a potential solution to this challenge.

As of 2020, almost every patient’s health information has been stored in some EHR system, and an increasing number of patients carry medical IoT devices, especially those with chronic diseases. However, we believe that the true value of medical devices has not been realized, given the minimal application of physiological data in clinical practices. Such a limited application could arise from various reasons, including conflicting incentives from different parties, imperfect regulations, technology barriers, etc. Our paper aims to provide a potential solution in the technical aspect of this issue. The foremost important issue is lack of interoperability between medical devices and clinical data infrastructures, such as the EHR system, where providers mainly store and retrieve patient health information. As a result, medical device data is not accessible by most providers when it could have been helpful to reach a proper clinical decision. Moreover, data security has always been a significant concern. For instance, storing data in a centralized system has been criticized to expose sensitive data to cyber-attacks. According to the Healthcare Data Breach report, millions of patient medical records are compromised every month [2]. Concerns on data security and patient privacy make the transmission of health information even more cumbersome and costly, which further impedes the improvement of healthcare decision making and the advancement of medical research.

To assuage these concerns, we propose an innovative information transfer system combining a blockchain-based technology and an API that enables interoperability between IoT medical devices and EHR systems, titled BlockIoT. First, BlockIoT serves as a method that allows medical devices to quickly and easily interface with the EHRs. It removes a significant amount of overhead (in the form of labor/development costs necessary to modify existing medical device protocols) necessary to transfer data between different parties that use individual proprietary systems to manage
data. The proposed system will allow individual medical device data to be securely transferred to providers without changing any firmware on medical devices or forcing companies to modify existing protocols when creating medical devices. As a result, it simplifies the required infrastructures for data transmission and reduces the number of middlemen involved, which could potentially improve cost efficiency in this process. Another advantage of BlockIoT is the ability to present providers with essential data points by communicating and analyzing medical device data. For instance, providers or insurance companies can use algorithms and visual diagrams to quickly understand medical data and determine whether the patient’s health is within normal limits or requires medical attention. This allows for quicker healthcare interventions, which prevents unnecessary and costly visits (such as to the emergency room) and allows patients’ symptoms to be resolved at a much faster pace than in the current healthcare setting. Notably, the proposed system can be combined with emerging technology such as AI/ML to improve the quality and efficiency of healthcare delivery. For example, it can be connected to an automatic alert system facilitated by AI technology to generate messages to providers about unexpected abnormal symptoms or to remind patients to comply with medication. Studies have shown that simple reminders or alerts in the form of text messages or emails can change patients’ behavior and improve their health as a result [3][4][5]. However, due to constraints in time and resources, such interventions from providers are often not readily available and thus, patients have to undertake most of the responsibility to stay compliant and follow the proper treatment regimen, which is particularly challenging for elderly patients with multiple complicated comorbidities.

We assess the proposed system by conducting structured interviews with healthcare practitioners from various backgrounds. During the interview, we first introduced BlockIoT and then presented a mockup of a sample EHR system integrated with BlockIoT. We also provided a History of Present Illness (HPI) description of a sample patient, who suffers from multiple complications and carries different types of medical devices. Then we asked providers opinions on how to treat the sample patient in the presence and absence of the device data. We also asked for suggestions on what additional information is desired and how the presentation of information can be improved. The majority of respondents viewed medical device data as enabling a more complete and objective picture of patients’ health. Moreover, the system can improve providers’ understanding of patients’ health and facilitate communication with patients. Adopting such a system can streamline the healthcare delivery workflow as long as the system is user-friendly and information can be presented clearly and succinctly. Our clinician evaluators, in general, acknowledge the value of the system, especially if the associated devices are also easy to use for patients.

Related Work

There have been several attempts at bridging blockchain, EHR, and wearable health IoT devices. However, our literature review indicated that the success of these systems had been limited. This section illustrates some of the comparable related works and how BlockIoT transcends the capabilities offered by them.

A system for remotely monitoring patients in a Health Insurance Portability and Accountability Act (HIPAA) compliant manner is introduced in [6][7]. However, the system does not focus on specific methods or protocols a device can use to communicate with the blockchain, forces device manufacturers to change device specifications to conform to the system’s requirements, no method is described that enables medical devices to integrate into the system, or in regards to representing this data to a physician. A data transfer system between a physician and patient through the use of a blockchain system for the transfer of labs and medical charts (rather than medical device data) is described in [8]. However, unlike in our system, this transfer of data is only one-sided and does not allow any analysis of the data present, which subjects the blockchain integration to be simply another form of a database. The integration of IoT devices to mobile applications is discussed in [9]. The framework manages medical devices but is very shallow in terms of its data transfer and storage. Primarily, representing data in a mobile application that is only available to a patient severely limits the potential use of the data itself. Secondly, because the system is centralized rather than decentralized, a significant risk is present in which any breach of that system can result in the release of significant amounts of medical data. Finally, there is no method present at all that can analyze or encrypt incoming data. An EHR system that can connect various IoT devices is introduced in [10]. However, the method it uses to achieve this has severe limitations. Data is primarily stored in a centralized rather than a decentralized system, which results in a significant risk in which any breach of that system can result in the release of significant amounts of medical data. Secondly, while data can be analyzed, the system can be integrated with only specific medical devices and support a
minimal scope of incoming data.

On improving EHR security, a system that enhances the security of current blockchains without the use of any specific keys is described in [11]. However, they do not provide practical methods to input or output data from the system. Instead, it forces a physician to manually input data into the system itself. Using blockchain to secure medical device data and prevent any malware or breaches from occurring is introduced in [12]. However, enforcement of this security comes with significant limitations as well. Primarily, no method is present to communicate medical data to another party, preventing physicians access to patient data. Secondly, similar to other papers above, the system forces manufacturers to create an entirely new device to satisfy the system’s needs, which prevents any commercial device to use the system.

On data sharing beyond the clinical settings, a secure method to transfer data between a researcher, patient, and physician through the use of blockchain is investigated in [13]. However, the paper does have some limitations. Primarily, no method or protocol exists for medical devices to send data in real-time to the system. Instead, the system leaves this data to the medical device company or the patients to do on their own. This process significantly discourages the transfer of medical device data from the patient to the relevant physician, and the data itself is unable to be analyzed in real-time, which prevents any life-threatening alerts from reaching the patient.

Methodology

The BlockIoT system is developed with four main functional goals. (1) Receiving real-time raw data from various commercial but standards-compliant medical devices; (2) converting raw data-points into well-defined and structured information, which are stored in a decentralized system; (3) analyzing raw data-points through smart contracts for potential patient or physician alerts; and (4) releasing this data to various EHR systems to represent medical data in a concise and easy-to-read fashion. These four aspects serve as a basis for the implemented work to increase patient data interoperability between medical devices and EHR systems. We first introduce the system’s architecture and the various parties involved and detail specific operations that highlight interoperability.

System Architecture

The BlockIoT system is constructed as a combination of an API and various smart contracts that connect existing EHR systems to several medical devices (Figure 1). It is primarily characterized as a decentralized system to store encrypted patient medical device data received from the API, which contains endpoints that receive data published by a patient’s medical IoT device. We define a medical IoT device to be an equipment that serves to collect and transmit patient health data from the patient’s location (such as the patient’s home) to a specific destination (such as commercial servers of the medical device company) using IoT. For example, a medical device can be a wearable heart-rate monitor that can transmit real-time heart rate to a specific destination over WiFi.

Figure 1: High-level Overview of the Data Streams Integrated into the BlockIoT System

Before describing the system’s implementation, some definitions are introduced first to establish the role of each component in the system.
**BlockIoT API** The API component is created to serve as a receiving endpoint for all incoming medical device data. Because of the immense number of medical devices found in the market today, the API was created with flexibility in mind. As a result, multiple communication protocols such as HTTP(S), MQ Telemetry Transport (MQTT) [14], and Constrained Application Protocol (CoAP) [15] are implemented in the API to ensure maximum possible coverage of all medical devices present in the market.

**Decentralized Storage** A decentralized network was chosen (rather than a centralized system) to be the most optimal solution to ensure security, immutability, and efficiency when storing the patient data. We implemented the storage for BlockIoT using the InterPlanetary File System (IPFS) along with the InterPlanetary Naming System (IPNS), in which each patient is connected through a peer-id that is generated based on their biometrics (first name, last name, and date of birth). IPFS was the primary choice of the BlockIoT system due to its ease of use and its ability to make data resilient over time. However, because the IPFS hashes change as the content of the file changes, IPNS is used to ensure that the address of patient content is the same regardless of the patient data change.

**System Features** The system is divided into two primary sections, including a medical device uploading data to the decentralized storage system, and an EHR requesting data regarding a certain patient.

**Security** Compared to more popular and widespread centralized systems, a blockchain network architecture is likely to be more secure due to the location of data storage. Because data is stored between trusted peers rather than in a central location, the risk of patient data being leaked in wholesale from a central location is very minimal. Patient data is stored through asymmetric key encryption, in which keys to the data belonging only to two parties, i.e., the physician’s EHR for a specific patient and the BlockIoT system, prevents an unauthorized user from accessing patient information. At the same time, patient data is only available for a specific amount of time before data access expires. This is achieved through smart contracts, which provide an IPNS link to the physician once the EHR system makes a request, and a transaction is created on the blockchain. After a certain period that is denoted on the contract, this link will expire, and another request must be made for extended access.

**Immutability** One of the main concerning aspects of centralized servers is that it is relatively easy to mutate incoming data, which can compromise the stored medical device data. The ledger in the underlying blockchain in the BlockIoT system solves this problem, as the data added to the distributed ledger is permanently stored, and any modification of the given data is discouraged, which ensures data integrity throughout BlockIoT.

**Storage** In medical devices, the amount of data collected about a patient can be generated as fast as once per second. This rate extended to a larger population of 10,000 patients can reach up to millions of different data points per minute, which is enough to overwhelm most, if not all, centralized servers over time. This issue can be alleviated through a blockchain, which is server-less and can distribute the storage capacity necessary for all of these data points throughout its users. Through this system, enormous amounts of data can be stored without significant risk of a crash.

**Template-driven Data Harmonization** Templates are methods that are used to analyze and identify incoming data. Because each medical device sends values with different keys, templates are essential in determining what type of data is transmitted from the device. Secondly, based on the device’s identifiers, one can verify the device based on its output itself. Each type of device will have its own template. For example, a heart rate sensor will have a template that will contain all types of information transmitted by the heart rate monitor (such as beats per minute and SpO2). The upper limit and lower limit values are also present to recognize whether a specific incoming value is within or outside a normal limit, leading to alerting the physician if the patient’s condition is life-threatening. These templates are essential in BlockIoT because not only are they used as a template for data storage, but they can also serve as labeled anonymous data that can serve as training data for AI/ML algorithms for that specific physiology.

While it is possible to send over raw medical device data to an EHR system, the data sent to the EHR system are not useful to the physician as is. It is better to facilitate the physician by providing them with vital statistics or charts. These summarized outputs enabled through smart contracts can quickly allow a physician to review the patient’s health metrics and decide further treatment as necessary. To facilitate this process, based on templates created for each type of data, physician-accessible graphs are generated and are exported to the EHR system. AI/ML algorithms are used to further determine trend lines and pinpoint metrics of interest. These algorithms can be trained to generate greater insight with real-time data, which may be further useful for a physician.
Making BlockIoT FHIR-Proof

As interoperability is a highly desired feature in BlockIoT and because we want to reach as many medical devices as possible, we designed BlockIoT to be compliant with existing standard EHR protocols that are widely used today. By being able to replicate and accommodate to specific standards that are used by a majority of the systems that are used in the healthcare industry, such as HTTP, one can state with a certain level of confidence that the proposed system will be able to function in a real-world scenario. However, we took an additional effort to make the BlockIoT system primarily compliant with the Fast Healthcare Interoperability of Resources (FHIR) Standard [16], a system of requirements and guidelines on EHR system development created by the standards organization HL7. We closely examined the HL7 FHIR documentation [16] and leveraged the concepts related to medical devices and observations to the BlockIoT system.

Current data in the BlockIoT system is stored in the IPFS system in a folder with various JSON files (depending on the number of medical devices that a patient has) containing patient data ordered by a timestamp. In this case, the file itself will serve as a patient resource designed to be interchangeably used between the BlockIoT system and the EHR system. Each resource contains a logical ID assigned upon the creation of the file. The resource is created as a JSON representation with various properties that describe the patient’s biometrics, the template used to identify the data, and the medical device data the patient has. A resource may also be an image that contains an interpretation of the data, such as a graph or specific data points a physician can examine before seeing the patient. The image resource is only created when an HTTP request is created to ensure that data is the most recent and relevant as possible.

Communication with a standardized EHR system is done through a RESTful API, which consists of various verbs used to transmit and receive resources (as defined above) from one system to another. Following this API protocol will ensure that transmittance is reliable and the incoming data is usable by a standard EHR system. The current system can conform to existing API requests properly as the routes present in the current system are identical to the routes described in the FHIR documentation. As a result, because the architecture of the resources and the syntax of the API conforms to the FHIR documentation, one can ensure with some confidence that the BlockIoT system will successfully communicate with an existing FHIR-based EHR system.

Blockchain and Smart Contracts

To implement smart contracts, and store transactions in a decentralized fashion, a blockchain system is implemented. In such an architecture, the nodes are the servers run by major third parties participating in this system, such as an EHR system at a hospital, manufacturer, and researchers who will employ smart contracts to send or receive patient data. Each type of node has a different level or type of access to data. EHR systems can only request data about a specific patient, for example. A smart contract will then be placed that allows physician access to the patient’s data representation (in the form of an IPNS link) for a 24 hour time period. After this time access is complete, the link will become non-existent, and another request would have to be made.

Smart contracts will then be used to analyze patient data and compare data with specific templates. For example, the “adherence” smart contract will be programmed to examine data and check for timestamp discrepancies, and therefore determine whether the patient has missed a dose of medication. Such smart contracts can also be used for more complex data types, such as EKG monitoring, where external or industry-standard algorithms can be used to examine EKG waves, and detect whether a heart arrhythmia is present. Through the use of smart contracts, passive data transmission can lead to more proactive interventions. Based on physician and patient preferences (which would be present in the patient’s configuration file), the smart contract can then send alerts to the relevant party if the algorithm detects an abnormality. For example, a physician can choose to ignore sharp increases in a patient’s blood sugar level if the patient had just had a meal. If this option is selected by the physician, the smart contract would first detect an increase in sugar levels, text the patient if he/she had just had a meal, and the patient’s response to determine whether an alert is necessary to send to the physician. Both physicians and patients can also choose when and how to receive alerts as well.

BlockIoT In Action

The BlockIoT model of data collection was evaluated with synthetic patient health data to test for practicality and usability. The data used for the system simulation was generated through Synthea [17], which is a patient population generator commonly used to produce synthetic data based on distributions learned from real patient data. We tested BlockIoT against 1000 generated patients, which is typical of the number of patients associated with a physi-
cian’s practice. These patient records contained different types of data, such as systolic/diastolic blood pressure, heart rate, body height/weight to represent basic vital signs, and more sophisticated lab results such as hemoglobin levels, leukocyte/erythrocyte counts.

First, the synthetically generated patients were used to determine whether uploading and downloading from IPFS is reliable, and a decentralized system can be used as a method to store medical device data. Out of the 1000 patients simulated, all 1000 patients were successfully securely uploaded to IPFS, and medical data was reliably stored. There was no missing data present, and no packet losses occurred when uploading data. Next, the BlockIoT model was evaluated with multiple types of medical data, as described above. Templates for each type of data were created, and the system was checked to see if any errors with data handling occurred. Out of the 8 different types of data tested (blood pressure, heart rate, blood oxygen, sugar levels, EKG, compliance, spirometry, and cell counts), the system handled all 8 types of data appropriately, in which, the data points that were not within normal limits were recognized and represented adequately in the physician user interface. Secondly, each data point was properly stored by timestamp, and the system also captured trends over time. Finally, the BlockIoT system was evaluated for the ability to receive a significant amount of incoming data at once. Because medical devices tend to send information extremely rapidly (in terms of seconds), the system must handle incoming requests quickly and efficiently to ensure that all data can be uploaded in real-time. 10,000 requests were to the system at a rate of 0.5 seconds for each request. The system was able to process each request within 5-6 seconds and handle all the incoming requests without delay.

**Participatory Design Survey**

Fourteen healthcare providers (thirteen physicians and one nurse practitioner) were invited to review the BlockIoT System in order to understand the need and impact of medical devices in a clinical setting, and determine whether there is potential in patient outcome improvement that can be attributed to integrating a system such as BlockIoT in their clinical practice. While we acknowledge that the sample size is not large enough to make a statistically significant conclusion about BlockIoT’s effectiveness, this evaluation aims to determine whether such a system has the potential to be feasible and useful in a participatory design setting for clinical data collection and usage scenarios. The providers were recruited through professional and personal relations of the researchers. However, a particular focus was put on inviting physicians from different specialties. As a result, the total sample of physicians consisted of a cardiologist, a gastroenterologist, a dermatologist, a geriatrician, an emergency physician, a surgeon, an anesthesiologist, a pulmonologist/allergy/sleep medicine physician, internal medicine specialist, and multiple primary care physicians. During the recruiting process, no prior information about the project was provided to the participants, including the concept, demonstration, or questions regarding BlockIoT. We followed this procedure to ensure authenticity during the interviews.

The recruited providers were given an option to either complete a form, or attend a virtual interview if they had time. The providers were not offered any remuneration for their answers. We conducted interviews through WebEx, and with permission from the providers, we recorded the sessions and later transcribed the discussions for analysis. For the forms, a Google Form (available at [http://bit.ly/blockiot-evaluation](http://bit.ly/blockiot-evaluation)) was used with a recorded demonstration of BlockIoT, descriptions and questions.

During the survey (for both survey types), each provider was shown a presentation describing the BlockIoT system. Afterward, the physician was shown a mockup of a sample EHR system with BlockIoT implemented. The mockup was designed to seem like an EHR system, which the physician would be most likely familiar with and therefore will easily visualize the integration. Each provider was also provided with a HPI of a sample patient. Table 1 summarizes the patient’s health status.

The sample patient described in the HPI was designed to have multiple complications, such as obstructive sleep apnea, asthma, high blood pressure, obesity, and diabetes. This detailed patient profile was compiled to accommodate each specialist who, in this case, will be treating that aspect of the patient (for example, a cardiologist will treat the patient’s blood pressure related issues). Secondly, different diseases tend to be associated with different types of medical devices. For example, a smart blood sugar monitor’s purpose and method of function are different from a smart blood pressure cuff. As a result, by understanding these different devices’ uses by different specialists, one can tailor the data representation through the specialty. Furthermore, the HPI was generated to be very realistic to a typical patient these
Table 1: Patient HPI Summary

<table>
<thead>
<tr>
<th>Patient Basics</th>
<th>Diagnoses</th>
<th>Devices</th>
<th>Current Medication</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wendy Barnes</td>
<td>Diabetes</td>
<td>Smart Blood Sugar Monitor</td>
<td>Metformin</td>
</tr>
<tr>
<td>(Age: 52; 80 bpm/98)</td>
<td>Obstructive Sleep Apnea</td>
<td>CPAP Machine</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>Hypertension</td>
<td>Smart Blood Pressure Cuff</td>
<td>Lisinopril</td>
</tr>
<tr>
<td></td>
<td>Asthma</td>
<td>Smart compliance tracker (for each medication)</td>
<td>Symbicort, Albuterol</td>
</tr>
<tr>
<td></td>
<td>Obesity</td>
<td>Smart weight tracker</td>
<td></td>
</tr>
</tbody>
</table>

providers would see in their clinical practice. While the patient is complicated, each diagnosis is associated with the other potentially co-morbid conditions. For example, increased risk of obesity is associated with high blood pressure, diabetes, asthma, and obstructive sleep apnea, which is further corroborated because typically, patients who have diabetes are more likely to develop cardiovascular disease. As a result, the other four diagnoses were also included in the HPI. Both correlations above were used to create the HPI of the sample patient. The provider was then shown 2-4 graphs representing sample medical device data (such as blood pressure readings, blood sugar levels, and compliance rates). The graphs shown to the provider depended on their specialty. For example, a pulmonologist would be shown a graph with the patient’s albuterol use and medication compliance in the last 30-60 days, and their blood pressure chart (Figure 2).

Figure 2: Synthetic Patient Charts from BlockIoT Shown to Evaluators

During each presentation, the provider was asked what difference was present between what is shown in the HPI and what the medical device data provides based on the information presented in the charts. We asked these questions to determine whether any significant impact is visible between the additional medical device data and whether this additional data can potentially change the provider’s medical decision. Secondly, the provider was also asked if the graph’s representation could be improved (such as changing the time axis or changing the type of graph present). This feedback can be used in future work, where the type, amount of data, or the time at which data arrives can help drive how the medical device information is represented to the provider. Moreover, each provider was asked about their overall review of the system and potential foreseeable drawbacks. At the end of the survey, for providers who were virtually interviewed, implications of the system in insurance claims, malpractice, medical data transfer from providers to providers, and potential impacts on the clinical workflow (in terms of efficiency) were also discussed.

We analyze the response data using a qualitative approach. In general, the providers’ responses were positive. All of the providers found the proposed system useful, with 87% of providers reporting that they would be willing to use the summary views as well as the patient/physician interventions throughout the patient care (the others were more inclined to just use the summary views of the medical device data). When asked whether the data coming from the BlockIoT system can change or influence the way the provider treats a patient, all the providers responded in agreement. Out of the providers asked about the efficiency impact of BlockIoT, all the providers expected the system
to streamline the workflow in the long run. Most of the providers stated that there are different levels of trust in the patient supplied information, and they use a variety of techniques to ascertain the information supplied to them by the patients during the visit. All the providers agreed that having longitudinal physiological data is better for the practice and makes them more efficient in the diagnosis process leading to better patient outcomes. However, at the moment, access to means for obtaining such data is minimal or non-existent. Physician 4 mentioned that they institute a log for patients with chronic illnesses; however, they admitted the quality of the data in these logs is only “about 50%”, which is not sufficient enough to make an accurate health suggestion. BlockIoT would be a trustworthy mechanism in which such data could be collected in a minimally intrusive way for the patients. Physicians 12 and 13 confirmed that having access to long-term blood pressure data for example, could help make them determine if a higher reading at the clinic is due to the “white coat hypertension” phenomena [13], especially if the readings taken at home as available through BlockIoT appear to be lower.

Confirming the main focus of BlockIoT, the need for more data-driven analyses in clinical practices was best summarized by Physician 3.

“The only way we know whether the patient has high blood pressure is whether the patient takes a one-time blood pressure reading a day at home in the morning, and that is only when we really ask to do it. Even then, they will only do it for some of the time, not all the time. Or when they come to the doctor’s office every 2 to 3 months. A blood pressure reading can change every second, which means that we are missing millions of data points right now!”

However, access to this device data is also variable, i.e., no standardized method exists currently. As best stated by Physician 1 when asked about retrieving data from various medical devices:

“It totally depends on the device; blood sugar monitors give you readouts, pacemakers need to be interrogated for data, blood pressure cuffs need to be brought to the office.”

Physician 2 also raised their concerns about the lack of integration of many of the devices patients use.

“For devices like the Apple Watch and Fitbit, there is no official way to get those reports. It’s essentially the patient showing you, like ‘this is what my Apple Watch said’ or ‘this is how much sleep I’ve been getting.’ For FDA-approved devices by Medtronic, Boston Scientific, those types of companies, the office will receive data from the companies, and it’s sort of like an additional provider level. But for other consumer devices, there is no specific support.”

To our question on AI-driven patient-centric notifications endorsed by the provider, all of our evaluators expressed enthusiasm. In particular, physician 2 said;

“It makes the patient-physician interaction more dynamic and more trustworthy, and the patient’s getting more involved with their care, and they’re more likely to be compliant with diet or lifestyle change or medication compliance if they know that they’re doctors following up with them.”

Physician 4 also mentioned that there would be some positive behavior change aspect when the patient knows that their data is being collected.

“It has to be easy for the patient because the patient knows ‘hey, somebody’s watching me. I can’t cheat. I can’t do it.’ This can make a huge difference.”

A specific interest the providers had was to not simply look at the analyses provided through the AI algorithms in their workflows but to also pinpoint certain aspects of the data between different time-frames, to look at anomalies presented in the data analysis closely, and interrogate the data. To this end, physician 2 said;

“We talked about having the ability to zoom in on certain times of day for certain parameters. I think that would be helpful, but it’s all going to come down to how easily I can access this information.”

However, some of the providers touched on the regulatory aspects of such a data-sharing ecosystem. In particular, physician 3 said;

“You have to make sure that when there’s a ton of data, there is always the patient’s approval or consent that is needed so that there is not an automatic transfer from my office to the hospital. You really have to have consent. In some countries, you will find it a bit tedious in terms of getting consent, while in other
health systems, it is going to be easy.”

The consensus on using the data supplied by BlockIoT for insurance claims, in authorizing tests, and in malpractice lawsuits had mixed reactions from the providers. Most of their concerns were centered around the fact that the current medical landscape is extremely complex, and any technological changes will have to be supplemented by regulatory interventions, and those changes will take some time to take effect.

**Conclusion**

Within the last decade, the rise of EHRs and medical devices has revolutionized the healthcare industry by altering how to capture and manage health information. Based on our participatory design survey, however, it appears that the information from an EHR is not nearly enough to result in an optimal clinical decision. For instance, there is no concrete method of determining or verifying how a patient’s health has evolved within a few months before a medical visit, as pointed out by most of our evaluators. Instead, most physicians determine diagnosis or set up treatment plans based on a snapshot of symptomatic information captured during the patient encounter or the patient’s subjective description, which presents a need to access more granular and real-time patient health information. As stated by all our evaluators whose specialty involves the usage of medical devices, data from these devices can be instrumental in determining compliance, detecting problems with patients whose disease has an asymptomatic nature (such as cardiac-related problems), or figuring out the change in patients’ health since the last visit. In traditional EHR systems, the accessibility to patients’ daily health data at a granular level is still extremely limited due to centralized and proprietary medical devices that cannot share patient data with third parties. In our survey, more than half of the evaluated physicians are unable to retrieve medical device data in their current clinical practice. Among those with limited access, they have to either ask patients to physically bring in the data or obtain the data via specific ties with the device companies.

To address these challenges, we propose and implement a universal, secure, decentralized bridge between patient-facing medical devices and physician-facing EHR systems, called **BlockIoT**. It is designed to provide real-time medical device data to physicians, who are usually constrained by their busy schedules, in an easy-to-understand manner so that they can use this data to provide higher-quality and more-efficient care. As per the evaluation, BlockIoT has shown potential in improving quality of care, patient-physician interaction, office visit efficiency, and overall more informed physicians’ decision-making. BlockIoT may also help physicians rule out certain diseases by offering additional data that underlie patients’ health to reduce time and costs that would have been spent on determining the diagnosis. Moreover, BlockIoT may also contribute to other aspects, such as justifying reimbursements or reducing malpractice suits due to its ability to provide objective and concrete patient data. A key advantage of BlockIoT is offering physicians an integrated look into data streams—that capture information from closely monitoring patient daily life and are traditionally unavailable to them—in a relatively easy and reliable way. Data from such an integrated system are more likely to form a complete and accurate health profile of patients, which could help improve the ML algorithm that is currently mostly based on biased, truncated or even distorted data. The data supplied to the EHR systems by BlockIoT contains more specific health information that could come from a variety of patients and can potentially be incorporated into a training algorithm without excessive efforts via the proposed system. The findings may assist physicians in making more viable treatment plans for the underrepresented population and even help reduce health disparity between groups. We believe that the proposed system, combined with the emerging technology like AI/ML, could have more far-reaching impacts. As pointed out by our physician evaluators, closely tracking patients’ health evolvement and offering timely interventions and reminders could be extremely valuable but is currently very challenging given the already-heavy workload. An automatic alert system facilitated by BlockIoT and supported by AI/ML technology could be transformative, let alone the potential feature of constantly-updated parameters thanks to the improving predictability of the training algorithm after incorporating the data from medical device.

The current version of BlockIoT, however, is far from perfect. Based on the providers’ interviews, the user interface and accessibility were the primary focus when asked about the potential drawbacks of the system. The consensus was that the system has the potential to be very useful to them in their clinical practice, as long as the data is easy to use and versatile enough to fit the specialist’s needs. A consent system is also necessary to ensure that patients can consent to what data should be shared with the physician and how long access should be present. After the system undergoes all possible upgrades, the system’s value needs to be further justified using a larger sample with sufficient statistical
power. Also, there remain significant challenges to persuade stakeholders to participate in the system due to varying incentives. While the proposed system is not a panacea for the issues plaguing the healthcare industry, we believe that our system establishes a building block for future research and enables a step forward in addressing all the challenges faced by healthcare providers in their clinical practice, ultimately leading to better patient outcomes.

Acknowledgements This work is partially supported by IBM Research AI through the AI Horizons Network and the Flash Grant from the School of Humanities, Arts, and Social Sciences (HASS) at Rensselaer Polytechnic Institute. Special thanks to the following physicians for providing their expert opinion: Dr. Anant Agarwalla, Dr. Neha Agarwalla, Dr. Vipin Agarwalla, Dr. Preeshini Fernando, Dr. Dara Fuentes, Dr. Danekka Loganathan, Dr. Maya Matsumoto-Bessam, Dr. Adriana Miranda, Dr. Arbol Ortiz, Dr. Bharati Reddy, Dr. Benjamín Reyna Sánchez, Dr. Mayank Shukla, Dr. Ransirini Wijeratne-Fernando and nurse practitioner Cynthia Jean-Baptiste.

References

Prediction of Resuscitation for Pediatric Sepsis from Data Available at Triage

Peter Stella¹, MD, MSc., Elizabeth Haines², MD, Yindalon Aphinyanaphongs³, MD

Affiliations:
1: Department of Pediatrics, NYU Grossman School of Medicine, New York
2: Department of Emergency Medicine, NYU Grossman School of Medicine, New York
3: Department of Population Health, New York University, New York

Abstract:

Pediatric sepsis imposes a significant burden of morbidity and mortality among children. While the speedy application of existing supportive care measures can substantially improve outcomes, further improvements in delivering that care require tools that go beyond recognizing sepsis and towards predicting its development. Machine learning techniques have great potential as predictive tools, but their application to pediatric sepsis has been stymied by several factors, particularly the relative rarity of its occurrence. We propose an alternate approach which focuses on predicting the provision of resuscitative care, rather than sepsis diagnoses or criteria themselves. Using three years of Emergency Department data from a large academic medical center, we developed a boosted tree model that predicts resuscitation within 6 hours of triage, and significantly outperforms existing rule-based sepsis alerts.

Introduction:

Estimates suggest that pediatric sepsis is responsible for approximately 75,000 hospitalizations and 10,000 deaths annually in the United States(1). While no specific treatment for sepsis exists, rapid volume repletion with intravenous fluids and the prompt administration of broad-spectrum antibiotics have been shown to decrease morbidity and mortality(2–5).

Current consensus guidelines emphasize that treatment should begin within one hour, and such rapid intervention requires rapid recognition, leading to a great deal of interest in systems to improve the accuracy, reliability and speed of sepsis diagnoses(6–13). The current guidelines suggest the use of a formalized screening tool to identify patients who meet sepsis criteria, as well as the use of a “bundle” of interventions to be performed should a positive screen occur(14). While these systems can be deployed as manual tools, they are well suited to integration into the electronic health record, and many institutions have done just this. But though useful and effective, this “screening” based approach has several intrinsic limitations.

The first is that current definitions of “sepsis” in pediatrics are extremely broad, particularly because pediatric definitions include the distinction between sepsis and “severe sepsis”, which has been abandoned in the adult criteria. In children, sepsis is defined as the presence of the “systemic inflammatory response syndrome” (“SIRS”) in the setting of known or suspected infection. SIRS itself is defined as a constellation of age adjusted vital sign abnormalities that must include temperature abnormalities or an abnormal white blood count. A patient who meets these sepsis criteria and who has evidence of organ dysfunction is said to have severe sepsis, and one who has sepsis and hypotension despite adequate fluid resuscitation is considered to have septic shock(15).

These definitions complicate attempts to build clinically relevant screening and prediction tools. Because SIRS vital sign criteria are generously defined (outside the 10th percentile for age), the category of “sepsis” can capture large numbers of children with mild, self-limited infections who are in little danger of progression towards organ dysfunction or death. Indeed, in one study SIRS criteria were found to be present in 15% of all emergency department visits, and over 80 percent of those children were discharged home without readmission(16).

This means that tools that identify “sepsis” often lack adequate specificity, with values around 85%. This can generate dozens of false positives a day in a busy emergency room, forcing hospitals systems to either de-tune their alerts to reduce false alarms or impose a significant burden on their clinicians with the concomitant risk of alarm fatigue. Furthermore, screening approaches fail to account for the fact that sepsis is a rapidly evolving process, and only approximately 40% of patients who develop severe sepsis meet criteria at presentation(6). Tools to identify
existing sepsis are undeniably helpful, but methods for predicting which patients will go on to develop sepsis prior to actually doing so have the potential to interrupt a dangerous and self-reinforcing process before injury is done.

Several groups have already published work using machine learning techniques to predict the occurrence of severe sepsis in selected populations, with good results(16,17). But while this work is important and likely to be clinically useful, it is not without drawbacks, because models that predict which patients will develop severe sepsis after presentation to an academic medical center effectively predict those patients who do so despite treatment, and therefore such models will not identify patients for whom resuscitative therapy interrupts physiological decompensation. We propose an alternative approach in which machine learning is used to predict which patients receive resuscitative treatment aimed at preventing the development of severe sepsis rather than its ultimate occurrence.

Altering the target of our predictive efforts from severe sepsis to resuscitation has several significant advantages. As above, it allows us to capture a group of patients who have the most to gain from more timely intervention- those for whom resuscitation can interrupt progression to severe sepsis. More prosaically, because initial resuscitation is a much more common event than the development of multi-organ dysfunction due to sepsis, it greatly increases the number of events on which we can train our algorithms.

This is not a trivial advantage. Because pediatric severe sepsis is rare, assembling a large enough cohort to derive meaningful predictive models requires multiple years of data from a large institution or a combination of data from multiple sources, the latter of which can create significant data-integrity problems. Furthermore, even if a predictive tool is developed using an enormous data set, it must be validated by potential users on “native” data before it can be responsibly employed, and then continuously surveilled lest its predictions “drift” off course. For any but the largest institutions, pediatric severe sepsis is simply too rare an event for this to be possible, whereas resuscitation occurs sufficiently frequently that validation and surveillance become feasible.

Methods:

We assembled data on 26,564 pediatric emergency medicine visits at New York University Langone Medical Center between January 1, 2017 and December 31, 2019. Data were extracted from the EHR data store (EPIC, Caboodle).

These data included demographics, triage vitals, triage nurse comments, chief complaint information, as well as orders placed and medications administered within 6 hours of arrival. Additionally, entries for the patient problem list were included if they were added prior to the date of arrival to prevent contamination. No laboratory testing data was included.

Visits were excluded from analysis if they were missing critical information including a weight associated with that visit (not necessarily recorded at triage), a date of birth, or any vital sign measurement at triage (HR, Temp, RR, SBP, DBP, O2 sat). A total of 2,760 visits (10%) were excluded, the vast majority (2,490) because of absent values for both systolic and diastolic blood pressures, leaving a final data set of 23,804 visits (see Table 1).

Table 1. Demographics

<table>
<thead>
<tr>
<th></th>
<th>Resuscitation</th>
<th>No Resuscitation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>Age</td>
<td>4.1</td>
<td>4.5</td>
</tr>
<tr>
<td>IV fluid Volume</td>
<td>29.0</td>
<td>13.2</td>
</tr>
<tr>
<td>Temp</td>
<td>100.6</td>
<td>2.0</td>
</tr>
<tr>
<td>HR</td>
<td>147</td>
<td>29.1</td>
</tr>
<tr>
<td>RR</td>
<td>33.2</td>
<td>13.2</td>
</tr>
<tr>
<td>O2 Saturation</td>
<td>96.1</td>
<td>5.1</td>
</tr>
<tr>
<td>SBP</td>
<td>115.6</td>
<td>17.5</td>
</tr>
<tr>
<td>DBP</td>
<td>72.4</td>
<td>17.1</td>
</tr>
<tr>
<td>Blood cultures</td>
<td>601</td>
<td></td>
</tr>
<tr>
<td>Antibiotics</td>
<td>601</td>
<td></td>
</tr>
<tr>
<td>Keyword + CC</td>
<td>354</td>
<td></td>
</tr>
<tr>
<td>Keyword + PL</td>
<td>108</td>
<td></td>
</tr>
<tr>
<td>Keyword + TC</td>
<td>62</td>
<td></td>
</tr>
</tbody>
</table>
Triage vital signs for use in the predictive models were normalized for age using median values for age, and no subsequent vital signs after triage were included in the data set. Vital signs for SIRS criteria were not transformed, but rather compared to age specific cut-offs specified by the SIRS criteria (See supplementary Table 1).

Membership in the resuscitation group was determined by the presence of order and medication administration data indicating that the child received:

1. >19 cc/kg of normal saline; AND
2. Broad spectrum parenteral antibiotics; AND
3. A blood culture,

all within 6 hours of admission to the Emergency Department.

The value of 19 cc/kg was chosen rather than 20 cc/kg because a large number of patients who were ordered to receive boluses of 20cc/kg were found to have received between 19-20cc/kg because of rounding related to common fluid bag sizes and pharmacy dispensation.

Text data, including the chief complaints, triage notes, and triage comments was tokenized and sorted based on word frequency. Words that occurred more than five times throughout the data set were reviewed, and any word related to known risk factors, including immunosuppression, malignancy, chronic disease, and the presence of indwelling lines, tracheostomies, and g/j- tubes, severe illness, or fever prior to presentation were added to a keyword list (see supplementary Table 2). The presence of any keywords meeting these criteria was recorded as a positive result for that category and each was encoded as a single separate feature (i.e. Problem list +, Chief complaint -, etc)

Models were fit on a training set composed of a random sample of 80 percent of the available data (n = 19,044) stratified to included 80 percent of the cases which met our resuscitation criteria, and then tuned using 5 fold cross validation with this data set. Final model performance was evaluated on a similarly stratified test set composed of the remaining 20 percent of the data (n= 4760), and this process was repeated 25 times to estimate confidence intervals. Several methods were employed, including standard and regularized regression (glm and elasticnet), random forests (randomForest, Rborist), gradient boosted trees (GBM and Xgboost), and generalized additive models (mgcv). All models were evaluated with the pROC and PRROC packages in R.

Results

As a baseline, we evaluated the performance of the existing SIRS criteria-based alert at triage to predict which patients would receive resuscitation. This alert performs relatively well with regard to specificity (0.909), but fired on fewer than half of the patients who received resuscitation within the next 6 hours (sensitivity 0.465). (See table 2).

Table 2. Confusion Table for SIRS Criteria Model and Resuscitation Outcome (Full Data set)

<table>
<thead>
<tr>
<th>SIRS model prediction</th>
<th>Did Not Receive Resuscitation</th>
<th>Received Resuscitation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative</td>
<td>21,093</td>
<td>321</td>
</tr>
<tr>
<td>Positive</td>
<td>2,110</td>
<td>280</td>
</tr>
<tr>
<td>Sensitivity</td>
<td></td>
<td>0.465</td>
</tr>
<tr>
<td>Specificity</td>
<td></td>
<td>0.909</td>
</tr>
<tr>
<td>PPV</td>
<td></td>
<td>0.117</td>
</tr>
</tbody>
</table>

We next built basic models using only the same data used by the SIRS criteria, with the addition of diastolic blood pressure and oxygen saturation levels. While this sort of sparse data is not ideal for machine learning applications, the value of using minimal data in model construction is that it allows for easy implementation in a real-world context as well as provides a “fair” comparison to the SIRS criteria, as little additional information is added. Performance improved slightly, with increases in the AUPRC to just over 0.15, with sensitivities set to match those of the SIRS criteria. (figure 1, table 3).
Table 3. Performance of models built on vital signs only for predicting resuscitation.

<table>
<thead>
<tr>
<th>Model</th>
<th>Sensitivity (95% C.I.)</th>
<th>Specificity (95% C.I.)</th>
<th>PPV (95% C.I.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIRS Criteria</td>
<td>0.437 (0.399-0.474)</td>
<td>0.910 (0.908-0.912)</td>
<td>0.116 (0.112-0.119)</td>
</tr>
<tr>
<td>LR Model</td>
<td>0.454 (0.443-0.468)</td>
<td>0.919 (0.913-0.924)</td>
<td>0.128 (0.122-0.135)</td>
</tr>
<tr>
<td>XGB Model</td>
<td>0.458 (0.443-0.472)</td>
<td>0.925 (0.919-0.930)</td>
<td>0.138 (0.131-0.146)</td>
</tr>
</tbody>
</table>

We next attempted to increase performance by adding additional data to our training set, in order to capture several known risk factors for severe sepsis using keyword searches of the patient’s chief complaint, triage comments, and problem list (prior to the day of presentation). These additional features improved model performance substantially, with AUPRCs rising from 0.15 to just under 0.20. The best performing model (Xgboost), was able to achieve a specificity of 94.3% and a positive predictive value of 17.6% while matching the sensitivity of the SIRS criteria (figure 2, table 4).

Table 4. Performance of models built on vitals + Keywords for predicting resuscitation

<table>
<thead>
<tr>
<th>Model</th>
<th>Sensitivity (95% C.I.)</th>
<th>Specificity (95% C.I.)</th>
<th>PPV (95% C.I.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIRS Criteria</td>
<td>0.437 (0.399-0.474)</td>
<td>0.910 (0.908-0.912)</td>
<td>0.116 (0.112-0.119)</td>
</tr>
<tr>
<td>XGB Model</td>
<td>0.455 (0.443-0.468)</td>
<td>0.943 (0.938-0.948)</td>
<td>0.176 (0.165-0.187)</td>
</tr>
</tbody>
</table>

Feature importance in the Xgboost model (figure 3) was analyzed by relative importance, showing very strong contributions from temperature, and significant contributions from chief complaint and problem list keywords, as well as most of other vital signs, although blood pressures, particularly diastolic blood pressures, had very little impact. Similar results were seen with other tree based techniques, and with regression based approaches as well.

Figure 1: Representative AUROC and AUPROC for models based on vital signs only
As an exploration of whether models built to predict resuscitation would perform well at predicting the development of severe sepsis, we assembled a list of 17 patients who were retrospectively judged (after manual chart review of administrative data) to have met criteria for severe sepsis within 6 hours of presentation to the emergency room. Of these patients, 11 met SIRS criteria at triage, and 15 were identified by our keyword trained models (a net increase of 4 patients).

Figure 2: Representative AUROC and AUPRC curves for models predicting resuscitation from data at triage using vitals and keywords.

Figure 3: Relative Variable importance for XGB Keyword Model
Discussion

Given the burden of morbidity and mortality caused by pediatric sepsis and the importance of speed in its treatment, there is a strong need for effective predictive risk models to allow clinicians to identify patients at high risk in order to intervene as quickly as possible. However, development of these models has been hampered by slippery definitions, the intrinsic rarity of severe sepsis, and the possible masking of cases by successful treatment. To mitigate these issues, we built models that attempt to predict resuscitative treatment rather than severe sepsis itself, and were able to significantly outperform existing alert systems, based on the SIRS criteria.

There are several weaknesses in this study, the most glaring of which is that while our ultimate goal is to predict those patients who should receive resuscitation in order to provide that information to clinicians, what we have done in this paper is predict those who will receive resuscitation. While logically flawed, we believe that this decision is defensible on several grounds. Most importantly, authoritatively answering who should receive treatment requires a randomized controlled trial of resuscitation criteria, which would be both resource intensive and ethically knotty. Additionally, while all physicians are capable of error, we believe that the pooled judgement of dozens of pediatric emergency physicians over three years provides a very close approximation of the standard of care, and therefore who should receive resuscitation.

That said, it is also true that the standard of care is necessarily imperfect, and this fact exposes another deep limitation in this approach. Because we base our predictions on clinical decisions that are already being made, we are unable to discover new risk factors or suggest a different approach where the current standard of care is systematically wrong. If we had performed this experiment in the middle ages, our models would confidently suggest bloodletting for a tertian ague. While we believe that using treatment decisions as targets can be very useful in developing tools for clinical decision support, it cannot replace more conventional definitions when trying to understand the phenomenology of sepsis itself.

The next area of difficulty is the definition of “resuscitation”. In developing our definition, we attempted to capture the first steps of resuscitative care based on national and institutional standards, but this definition remains arbitrary. This can clearly be seen from our decision to reduce our fluid requirement from 20cc/kg to 19cc/kg when inspection of the data revealed a large number of children in the 19-20cc/kg range due to rounding and commonly available fluid bag sizes. While this decision is certainly defensible, a similar argument could be made for a cutoff of 18 or 17.5 cc/kg, but ultimately any computational definition of what qualifies as resuscitation is a balance between a “strict” definition that misses instances of resuscitation, and a very “loose” one that introduces noise by capturing patients who received these treatments for reasons other than a risk of severe sepsis.

This latter effect is clearly visible in our data. During subjective validation of our results, we found numerous examples of younger children with known, but non-life-threatening infections such as cellulitis or oro-facial infections who were judged to have been “resuscitated” by virtue of receiving maintenance or “maintenance and a half” fluids throughout the first 6 hours of their ER stay.

Another notable limitation of this study is that we have only examined data from a single point in time. The decision to focus on the moment of triage is based on the desire to develop a system that provides the earliest warning possible based on the least amount of data, but it is clear that in doing so we are excluding highly valuable information, particularly that from laboratory testing and longitudinal changes in vital signs. It is not uncommon for children to present with very early sepsis, and for resuscitation to begin after the child has “declared” himself while in the emergency room. This is simply the nature of the problem, and other authors have reported only approximately 40% of patients who go on to develop severe sepsis meet criteria at presentation. Indeed, review of model errors showed that patients who are afebrile at presentation are almost certain to be missed by both machine learning and criteria-based models. Models like ours, which are designed to operate at triage, must be supplemented by additional tools that can perform continued surveillance while taking advantages of the richer data streams that accumulate as a patient’s stay progresses.
Conclusion

Despite these limitations, we believe that there are two useful conclusions to be drawn from this data. The first is not strictly related to our project of developing machine learning models for resuscitation prediction, but relevant nonetheless: the SIRS criteria, when applied at triage, do not perform particularly well at identifying children who will receive resuscitation in the near term. Indeed, a SIRS based alert system fires on fewer than half of the total patients who go on to receive resuscitation within 6 hours, while generating more than 7.5 false positives for each patient who does receive treatment. These results are generally in line with other studies that have used formal definitions of sepsis (ICD codes and manual chart review), and further underscore the contention that these SIRS criteria alone are not well suited to implementation as part of clinical decision support, which requires that the information relayed to providers be of high enough quality to meaningfully alter their treatment decisions.

The second conclusion is that training models to predict resuscitation, rather than severe sepsis, can provide a way to generate actionable decision support. While we believe that use of criteria confirmed severe sepsis as a target is flawed because it fails to account for “interrupted” cases of sepsis, the most fundamental advantage of focusing on resuscitation is simply that many more resuscitation events occur than instances of severe sepsis. By increasing the number of events, this “looser” target ameliorates one of the core problems in applying machine learning to pediatric sepsis, particularly the difficulty of fitting and validating models for very rare events.

Another important advantage of focusing on resuscitation as a predictive target is that aligns very closely with the real-world issues facing clinicians. In an emergency setting, the most immediate question is whether to begin initial resuscitation and workup, and by using this as a target, we generate a prediction that is easy to operationalize in a by linking the prediction of these models to the EHR in order to provide decision support.

But when considering the application of these models to decision support, several additional factors must be considered. While we achieved the best performance by using an ensemble of boosted trees (Xgboost), these models are relatively difficult to integrate into EHR systems, and have the further downside of limited explainability. This can both reduce clinicians’ trust in the model’s output and make it more difficult for them to identify situations where the model is inapplicable. Use of alternative model types, such as the use of explainable boosted trees and generalized additive models can provide much more comprehensible predictions, with only moderate loss of performance. As an example, a generalized additive model fit to our data provided a positive predictive value of 18% compared to Xgboost’s 20%, but can provide straightforward explanation of how each variable contributes to an individual prediction.

This is particularly important when decision support is being offered in a critical situation, as would be the case when a clinician is considering whether or not to start resuscitation out of a concern for sepsis. Very few practitioners or institutions are likely to be comfortable trusting the predictions of a mysterious algorithm when the stakes are so high, and model selection should be driven by a balance of ease of implementation, interpretability, and performance.

Regardless of the model chosen, great caution should be taken when considering the use of predictive algorithms to provide clinical decision support. While we believe that we used appropriate methods to limit the risk of overfitting, and report only test set results, it is well known that machine learning methods generally perform more poorly in the real world than on the data from which they were built, and any potential application will require thorough prospective validation before use.
References:


Supplementary Tables

<table>
<thead>
<tr>
<th>SIRS Triggers</th>
<th>Median Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (Years)</td>
<td>HR</td>
</tr>
<tr>
<td>&lt;1</td>
<td>180</td>
</tr>
<tr>
<td>1</td>
<td>180</td>
</tr>
<tr>
<td>2</td>
<td>140</td>
</tr>
<tr>
<td>3</td>
<td>140</td>
</tr>
<tr>
<td>4</td>
<td>140</td>
</tr>
<tr>
<td>5</td>
<td>140</td>
</tr>
<tr>
<td>6</td>
<td>130</td>
</tr>
<tr>
<td>7</td>
<td>130</td>
</tr>
<tr>
<td>8</td>
<td>130</td>
</tr>
<tr>
<td>9</td>
<td>130</td>
</tr>
<tr>
<td>10</td>
<td>130</td>
</tr>
<tr>
<td>11</td>
<td>130</td>
</tr>
<tr>
<td>12</td>
<td>130</td>
</tr>
<tr>
<td>13</td>
<td>110</td>
</tr>
<tr>
<td>14</td>
<td>110</td>
</tr>
<tr>
<td>15</td>
<td>110</td>
</tr>
<tr>
<td>16</td>
<td>110</td>
</tr>
<tr>
<td>17</td>
<td>110</td>
</tr>
<tr>
<td>Broad Spectrum Antibiotics</td>
<td>Problem List Keywords</td>
</tr>
<tr>
<td>---------------------------</td>
<td>-----------------------</td>
</tr>
<tr>
<td>Amoxicillin/Sulbactam</td>
<td>Chemo</td>
</tr>
<tr>
<td>Ceftriaxone</td>
<td>Leukemia</td>
</tr>
<tr>
<td>Cefepime</td>
<td>Cancer</td>
</tr>
<tr>
<td>Cefexime</td>
<td>Rhabdomyosarcoma</td>
</tr>
<tr>
<td>Cefotaxime</td>
<td>Neutropenia</td>
</tr>
<tr>
<td>Ertapenem</td>
<td>Thrombocytopenia</td>
</tr>
<tr>
<td>Gentamicin</td>
<td>Medulloblastoma</td>
</tr>
<tr>
<td>Linezolid</td>
<td>Mass</td>
</tr>
<tr>
<td>Meropenem</td>
<td>Sickle</td>
</tr>
<tr>
<td>Piperacillin/Tazobactam</td>
<td>Mucositis</td>
</tr>
<tr>
<td>Vancomycin</td>
<td>Immunocompromised</td>
</tr>
<tr>
<td></td>
<td>Immunodeficiency</td>
</tr>
<tr>
<td></td>
<td>Epilepsy</td>
</tr>
<tr>
<td></td>
<td>Delayed</td>
</tr>
<tr>
<td></td>
<td>Cerebral Palsy</td>
</tr>
<tr>
<td></td>
<td>Jejeunostomy</td>
</tr>
<tr>
<td></td>
<td>Gastrostomy</td>
</tr>
<tr>
<td></td>
<td>Feeding Tube</td>
</tr>
<tr>
<td></td>
<td>Clostridium</td>
</tr>
</tbody>
</table>
Assessing the Readability of App Descriptions and Investigating its Role in the Choice of mHealth Apps: Retrospective and Prospective Analyses

Wu-Chen Su, MS1, Khyati Y. Mehta, MS1, Kirandeep Gill, MS1, Peng Yeh, MS3, Ming-Yuan Chih, PhD4, Danny T.Y. Wu, PhD, MS12
1Department of Biomedical Informatics, University of Cincinnati; 2Department of Pediatrics, University of Cincinnati, Cincinnati, OH; 3Department of Statistics, University of Kentucky; 4College of Health Sciences, University of Kentucky, Lexington, KY

Abstract

People with low health literacy are more likely to use mobile apps for health information. The choice of mHealth apps can affect health behaviors and outcomes. However, app descriptions may not be very readable to the target users, which can negatively impact app adoption and utilization. In this study, we assessed the readability of mHealth app descriptions and explored the relationship between description readability and other app metadata, as well as description writing styles. The results showed that app descriptions were at eleventh- to fifteenth-grade level, with only 6% of them meeting the readability recommendation (third- to seventh-grade level). The description readability played a vital role in predicting app installs when an app had no reviews. The content analysis showed copy-paste behaviors and identified two potential causes for low readability. More work is needed to improve the readability of app descriptions and optimize mHealth app adoption and utilization.

Introduction

Health literacy is a critical factor affecting health behaviors and decisions and further impact health outcomes1. The readability level of patient education materials (PEMs), however, may not match the health literacy of the general public2. Additionally, studies have reported that 35% of adults in the United States (US) had basic or below basic health literacy3 and 47% of European Union adult residents had limited health literacy4. To meet the needs of average Americans who read at the eighth-grade level5, the American Medical Association (AMA) and the National Library of Medicine at the National Institutes of Health (NIH/NLM) have suggested that PEMs be composed on a third to eighth-grade level, or even fifth to sixth grade level6,7. PEMs failing to meet this required reading level (RRL) may not effectively delivery health information to their readers.

With the advances of information technology, PEMs have been published online and widely accessed by lay persons. Overall, 63% of US users and 71% of European users search health information on the Internet for a variety of purposes8-10. People with low self-reported health literacy are more likely to use mobile apps and to get health information from social networking sites11. However, this population may have limited ability to read app descriptions and understand app purposes, which can negatively impact their app choices. Meanwhile, the large amount of healthcare mobile apps (“mHealth apps” hereafter) available on the markets imposes challenges to the choice of suitable apps. These mHealth apps may target multiple user groups, including students, patients, healthcare professionals, and policy-makers12,13, which adds to the challenge of app choices. Recent studies have explored the factors behind app choices and found that users from the US are more likely to download medical apps, and that price, app features, descriptions, reviews, and rating stars were the most important factors affecting app choices14,15. These factors can also be confounded with each other. For example, paid apps (price) are expected to provide better quality than free apps in most regards, and thus a less readable description may diminish the perceived value of paid apps16. While many studies have enhanced the understanding of these decision factors and their relationships, no study so far focuses on improving the readability of app descriptions to help app developers and vendors design a better app and improve app utilization.

In this study, we aimed to assess the readability of app descriptions and explore the role of readability in app choices. Readability assessment has been conducted on online health information17,18 as well as other types of health information such as electronic health records. For example, a recent retrospective study revealed that free-text directions on electronic prescriptions (information about medications) can be less readable to the patients19. This study found that 51.4% of randomly sampled directions from 966 patients have at least one quality issue, and that the pharmacy staff had to transcribe these directions to make them more readable. Similar to the issue of understanding the directions of medication, many mHealth apps may confuse their potential users if the readability of app descriptions is low.
Here we had three research questions. Firstly, what were the RRL of app descriptions and their relationships with other app metadata in the major app markets, namely the Apple App Store and Google Play Store? Also, did the app readability meet the recommended readability guideline by AMA and NIH/NLM? We hypothesized that the app readability would not meet the recommended readability level similar to many online health information. Secondly, did the app readability play any role in app choices? The literature showed multiple factors, such as number of reviews and ratings, can play a significant role in app choices. We hypothesized that users were more likely to download and/or install an app if the RRL of app description is low (i.e., a lower readability score can predict a higher number of app installs). Finally, what would be the writing styles and mechanics contributing to the higher RRL of app descriptions? What were possible reasons for app developers to write a less readable description? Had any app developers taken any actions to improve the description readability? By answering these research questions, we aimed to bridge the knowledge gap, help increase the readability of app descriptions, and further assist the lay public to choose proper mHealth apps to improve their health.

Methods

Data Collection

In order to conduct large-scale app analyses, a request was sent to the authors who created a mHealth app repository\(^1\). This mHealth app repository contained rich metadata of ‘Medical’ (ME) and ‘Health and Fitness’ (HF) apps on Apple App Store and Google Play Store, leading to four subsets of data based on the two app categories and the two vendor markets. Table 1 shows the metadata collected in each of the markets. This repository was released in four time periods, namely, the second and fourth quarters in 2015 and 2016 (Q2/2015, Q4/2015, Q2/2016 and Q4/2016). In the present study, only apps with English descriptions and published on the US market were selected. In order to identify apps with English-only descriptions in the repository, a Python library called “langdetect” was used\(^2\). The metadata of the selected mHealth apps were processed and stored in our MySQL database.

<table>
<thead>
<tr>
<th>Vendor Market</th>
<th>Metadata List</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apple App Store</td>
<td>unique app id, developer id, average of user ratings, content rating, app description, price, category, number of user ratings.</td>
</tr>
<tr>
<td>Google Play Store</td>
<td>unique app id, average of user ratings, category, content rating, app description, developer id, number of installs, price, number of user ratings, video URL, number of screenshots, age of app (the difference between app release date, and data recorded date), the average ratings of top 4 most helpful reviews, the average ratings of top 10 most helpful reviews, the average ratings of top 10 recent reviews.</td>
</tr>
</tbody>
</table>

Language Surface Metrics

The app description was characterized using surface metrics including average document length (ADL) and vocabulary coverage (VC) as listed in Table 2. A general English dictionary, GNU Aspell\(^3\), was used to compute the VC of app descriptions and has been used in previous studies\(^4,5\).

Readability Measures

Two readability measures were utilized to assess the descriptions of healthcare apps, including Flesch-Kincaid Grade Level (FKGL)\(^6\) and Gunning-Fog Index (GFI)\(^7\), which were chosen in the present study due to their simple calculation and easy interpretation of the scores. These readability measures are considered “classic” measures due to their general design. They have been widely used since the 1970s and recently been used to assess online PEMs. Microsoft Word has implemented FKGL to allow users to assess and report the readability level of a document\(^8\). To simplify the interpretation of RRL, an average readability score (ARS) of FKGL and GFI was used since both measures are strongly correlated\(^2,9\). Table 2 lists the definition of the readability measures.
over time. The or just copy was investigated "copy and paste" behaviors using a python library called 'FuzzyWuzzy'.

SUMMARIZED:

Lastly, a content analysis was conducted to understand the description writing styles. Of note, analyzing the concepts of the app descriptions was out of the scope of the present study. The first step in this analysis was to summarize the number of apps that each developer created in each market, assuming that a larger portion of apps were developed by a small group of app developers who created multiple apps. The second step followed the first and investigated “copy and paste” behaviors using a python library called ‘FuzzyWuzzy’. A percentage was calculated by averaging the pairwise text similarity of the app descriptions of each app developer. The rationale was that copy-paste behaviors can exist if a developer creates multiple apps and uses a template to write the descriptions or just reuse existing descriptions. In this case, the readability of app descriptions would not change dramatically over time. The last step of the content analysis was to randomly select 50 apps that had a description with readability

<table>
<thead>
<tr>
<th>Surface Metrics and Readability Measures</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Document Length by Sentence (ADL-SENT)</td>
<td>Average number of sentences per document</td>
</tr>
<tr>
<td>Average Document Length by Lexicon (ADL-LEX)</td>
<td>Average number of lexicons per document</td>
</tr>
<tr>
<td>Vocabulary Coverage (VC)</td>
<td>Number of words covered by a dictionary normalized by the vocabulary size (percentage)</td>
</tr>
<tr>
<td>Flesch-Kincaid Grade Level (FKGL)</td>
<td>(0.39 \times \frac{\text{number of lexicons}}{\text{number of sentences}} + 11.8 \times \frac{\text{number of syllables}}{\text{number of lexicons}} - 15.59); a raw score can be rounded to the nearest integer below its current value to form a grade level.</td>
</tr>
<tr>
<td>Gunning-Fog Index (GFI)</td>
<td>(0.40 \times \left(\frac{\text{number of lexicons}}{\text{number of sentences}} + 100\right) \times \frac{\text{number of difficult words}}{\text{total lexicons}}); a word is considered difficulty if it has 3 or more syllables.</td>
</tr>
<tr>
<td>Average Readability Score (ARS)</td>
<td>The average of the raw scores of FKGL and GFI. Similarly, an ARS score can be rounded to the nearest integer below its current value to form a grade level.</td>
</tr>
</tbody>
</table>

Data Analysis

The data analysis included three components; each corresponding one of the three research questions. In the first analysis, a retrospective analysis was conducted to describe the mHealth apps statistically and compare the RRL among the groups using the surface metrics and the readability measures listed in Table 2. Specifically, the distribution of mHealth app groups by category, vendor market, and over four time periods were summarized. The ADL-SENT, ADL-LEX, VC were applied to characterize the app descriptions. Non-parametric (Mann-Whitney U) tests were used with the significance level of 0.05 to examine if there was any significant difference among the group medians.

Specifically, a total of 240 free Google Play apps listed in the mHealth app repository were crawled for three months (mid-April to mid-July of 2019) using a Python library called ‘play-scraper’. The actual number of apps included in this analysis was determined by the alive apps, the capacity of the Python library, and the study period. The included apps, regardless of their categories (i.e., ME vs HF), were further separated into two groups based on their number of reviews. The apps without any review were put into one group while the apps with at least one review were put into the other group. Linear regression in the R statistical package was used to model the data in each group, with the number of installs as the dependent variable and key metadata as the independent variable (predictors). The selection of metadata (independent variables) were based on the literature, including but not limited to: number of reviews, number of review rating, number of screenshots, and readability scores. The coefficient of determination (R-squared value) of each mode was reported to show the proportion of variance in the dependent variable explained by the independent variables.

Finally, a content analysis was conducted to understand the description writing styles. Of note, analyzing the concepts of the app descriptions was out of the scope of the present study. The first step in this analysis was to summarize the number of apps that each developer created in each market, assuming that a larger portion of apps were developed by a small group of app developers who created multiple apps. The second step followed the first and investigated “copy and paste” behaviors using a python library called ‘FuzzyWuzzy’. A percentage was calculated by averaging the pairwise text similarity of the app descriptions of each app developer. The rationale was that copy-paste behaviors can exist if a developer creates multiple apps and uses a template to write the descriptions or just reuse existing descriptions. In this case, the readability of app descriptions would not change dramatically over time. The last step of the content analysis was to randomly select 50 apps that had a description with readability

**Table 2. Definition of surface metrics and readability measures**

<table>
<thead>
<tr>
<th>Surface Metrics and Readability Measures</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Document Length by Sentence (ADL-SENT)</td>
<td>Average number of sentences per document</td>
</tr>
<tr>
<td>Average Document Length by Lexicon (ADL-LEX)</td>
<td>Average number of lexicons per document</td>
</tr>
<tr>
<td>Vocabulary Coverage (VC)</td>
<td>Number of words covered by a dictionary normalized by the vocabulary size (percentage)</td>
</tr>
<tr>
<td>Flesch-Kincaid Grade Level (FKGL)</td>
<td>(0.39 \times \frac{\text{number of lexicons}}{\text{number of sentences}} + 11.8 \times \frac{\text{number of syllables}}{\text{number of lexicons}} - 15.59); a raw score can be rounded to the nearest integer below its current value to form a grade level.</td>
</tr>
<tr>
<td>Gunning-Fog Index (GFI)</td>
<td>(0.40 \times \left(\frac{\text{number of lexicons}}{\text{number of sentences}} + 100\right) \times \frac{\text{number of difficult words}}{\text{total lexicons}}); a word is considered difficulty if it has 3 or more syllables.</td>
</tr>
<tr>
<td>Average Readability Score (ARS)</td>
<td>The average of the raw scores of FKGL and GFI. Similarly, an ARS score can be rounded to the nearest integer below its current value to form a grade level.</td>
</tr>
</tbody>
</table>
at the eighth-grade level and above in each of the market, totaling 100 apps. These 100 app descriptions were manually reviewed by the research team to form suggestions.

Results

Descriptive Statistics and Group Comparisons

Figure 1 shows the numbers of mHealth apps based on their attributes in vendor markets (Apple App Store and Google Play Store), category (ME and HF), and published time period (Q2/15, Q4/15, Q2/16, and Q4/16). Each group for comparison was denoted using the abbreviation of the attributes. For example, Health and Fitness apps on the Apple App Store were denoted as “A_HF”. Similarly, Medical apps on Google Play Store were denoted as “G_ME”. Here bar charts instead of line charts were used to show the comparisons because of our focus on vendor market and app category.

Overall, Apple App Store has more apps than Google Play Store. The average quarterly increase rate of the numbers of apps was more than 10% between 2015 and 2016 for apps in the Apple App store. For Google Play store, there was a similar tendency, although the average increase rate is more than 45%. In both markets, HF apps contributed more to the increase than ME apps; G_HF apps outgrew A_ME apps in Q4 of 2016.

![Figure 1. Distribution of the selected mHealth Apps.](image)

Table 3 shows the summary of surface metrics. Apps on the Google Play Store had more sentences and lexicons in their descriptions than those on the Apple App Store. However, there is no statistical difference of the average document length (ADL-LEX and ADL-SENT) among the app categories between the markets. The app descriptions were around 200 words in 10 sentences. The VC is high (>90%) in both markets, meaning the app descriptions were written in common vocabularies most of time rather than rare terms.

<table>
<thead>
<tr>
<th>Apple App Store</th>
<th>Health and Fitness (A_HF)</th>
<th>Medical (A_ME)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Study Period</strong></td>
<td><em><em>LEX</em> (std)</em>*</td>
<td><em><em>SENT</em> (std)</em>*</td>
</tr>
<tr>
<td>Q2/15</td>
<td>188.56 (150.96)</td>
<td>9.88 (8.24)</td>
</tr>
<tr>
<td>Q4/15</td>
<td>185.28 (149.04)</td>
<td>9.65 (8.06)</td>
</tr>
<tr>
<td>Q2/16</td>
<td>179.14 (147.87)</td>
<td>9.27 (7.87)</td>
</tr>
<tr>
<td>Q4/16</td>
<td>173.05 (145.79)</td>
<td>8.99 (7.79)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Google Play Store</th>
<th>Health and Fitness (G_HF)</th>
<th>Medical (G_ME)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Study Period</strong></td>
<td><em><em>LEX</em> (std)</em>*</td>
<td><em><em>SENT</em> (std)</em>*</td>
</tr>
<tr>
<td>Q2/15</td>
<td>224.81 (166.67)</td>
<td>11.83 (9.81)</td>
</tr>
<tr>
<td>Q4/15</td>
<td>190.53 (155.10)</td>
<td>10.18 (8.78)</td>
</tr>
<tr>
<td>Q2/16</td>
<td>185.32 (158.80)</td>
<td>9.72 (8.6)</td>
</tr>
<tr>
<td>Q4/16</td>
<td>184.82 (164.57)</td>
<td>9.67 (8.76)</td>
</tr>
</tbody>
</table>

*LEX: average number of lexicons per document. SENT: average number of sentences per document.

bstd: standard deviation.
Figure 2 shows that the ARS for these four groups of apps was between eleventh- and fifteenth-grade level. They were far beyond the recommended readability levels from AMA and NIH/NLM. The RRLs of app descriptions on the Apple App Store are statistically lower than these on the Google Play Store combining the categories and the time periods (Mann-Whitney U test, p < .01). Further, the descriptions of ME apps had higher RRL than HF apps in both markets (Mann-Whitney U test, p < .01). Additionally, only around 6% of apps on average on both markets met the readability recommendation, where more HF apps met the readability recommendation than ME apps (Figure 3).

![Figure 2. Average readability scores (ARS) for different groups and time periods.](image1)

![Figure 3. Percentage of Apps meeting recommended readability guideline (third- to seventh-grade level).](image2)

Next, the relationships between the RRL and the two metadata (i.e., price and content rating) were examined to understand the potential impact of description readability on app users. The analysis indicated that the paid apps did not have a lower RRL than the free apps. Meanwhile, suitable content may not always be readable to the targeted app users, which may have a negative impact on app choices and utilization. Specifically, apps with ‘4+’ content rating on the Apple App Store and ‘Everyone’ content rating in the Google Play Store were selected for analysis. As shown in Table 4, these apps categorized for any age group required their users with at least eleventh-grade level literacy to read the descriptions.

<table>
<thead>
<tr>
<th>Group</th>
<th>Age groups</th>
<th>A_ME</th>
<th>A_HF</th>
<th>G_ME</th>
<th>G_HF</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>4+</td>
<td>4+</td>
<td>Everyone</td>
<td>Everyone</td>
</tr>
<tr>
<td>Q2/15</td>
<td>14.07 (6.14)*</td>
<td>11.62 (6.52)</td>
<td>14.05 (6.86)</td>
<td>11.9 (6.4)</td>
<td></td>
</tr>
<tr>
<td>Q2/16</td>
<td>14.05 (6.36)</td>
<td>11.48 (6.18)</td>
<td>14.33 (6.68)</td>
<td>12.07 (7.78)</td>
<td></td>
</tr>
<tr>
<td>Q4/16</td>
<td>14.02 (6.34)</td>
<td>11.33 (5.94)</td>
<td>14.31 (7.11)</td>
<td>12.04 (7.02)</td>
<td></td>
</tr>
</tbody>
</table>

* Medical apps on the Apple App Store with ‘4+’ content rating in the second quarter of 2015 had a mean readability score (ARS) of 14.06 with a standard deviation of 6.14.

### Model Readability and App Installs on Google Play Store

When we tracked apps prospectively, the total number of apps dropped from 240 (collected in the mHealth app repository) in mid-April to 218 (alive in the app stores when crawling) in mid-July. The modeling results (Table 5) show that the free Google Play apps with no review had three significant independent variables (p < .01) to predict the number of installs. These significant independent variables (predictors) include the number of user ratings (β = 60.08), the number of screenshots (β = -6.93), and the average readability score (ARS, β = -4.62), with an intercept of 140.64. However, the effect size of this model was weak (R-squared=0.203). On the other hand, the free Google play apps with at least one review had six predictors significant (p < .01), including the number of user ratings (β =
39.52), the average user rating ($\beta = -1308.85$), the number of screenshots ($\beta = -142.35$), the average rating of top 4 most helpful reviews showing on the landing page ($\beta = 1517.23$), the average rating of top 10 most useful reviews ($\beta = -9544.52$), and the average rating of top 10 most recent reviews ($\beta = 8511.05$, with an intercept of 10114.91. The effect size of this model was moderate (R-squared=0.5332). The number of user ratings and screenshots were significant in both groups with and without a review, which speaks to the universal importance of these variables. When there was no review at all, the average readability score of the app descriptions seemed to play a significant role in predicting the number of installs of an app. The negative coefficient supported our hypothesis: the lower a RRL (ARS) is, the more the number of app installs would be. On the other hand, when there was at least one review, the predictors were dominated by the review ratings.

**Table 5. Factors contributing to the number of app installs on Google Play Store.**

<table>
<thead>
<tr>
<th>Coefficients</th>
<th>App with no review</th>
<th>p-value</th>
<th>App with at least one review</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>140.64</td>
<td>4.67e-6 ***</td>
<td>10114.91</td>
<td>7.61e-16 ***</td>
</tr>
<tr>
<td>Number of user ratings</td>
<td>60.08</td>
<td>&lt; 2e-16 ***</td>
<td>39.52</td>
<td>&lt; 2e-16 ***</td>
</tr>
<tr>
<td>Average user rating</td>
<td>7.14</td>
<td>0.0933</td>
<td>-1308.85</td>
<td>0.0014 **</td>
</tr>
<tr>
<td>Has video or not</td>
<td>-16.21</td>
<td>0.7066</td>
<td>-1012.84</td>
<td>0.1215</td>
</tr>
<tr>
<td>Number of screenshots</td>
<td>-6.93</td>
<td>4.06e05 ***</td>
<td>-142.35</td>
<td>4.28e-16 ***</td>
</tr>
<tr>
<td>App age (days on market)</td>
<td>0.22</td>
<td>0.7019</td>
<td>5.39</td>
<td>0.7268</td>
</tr>
<tr>
<td>Average readability score (ARS)</td>
<td>-4.62</td>
<td>0.0021 **</td>
<td>-5.33</td>
<td>0.5807</td>
</tr>
<tr>
<td>Number of reviews</td>
<td>NA</td>
<td>NA</td>
<td>-6.94</td>
<td>0.0635</td>
</tr>
<tr>
<td>Average rating of all reviews</td>
<td>NA</td>
<td>NA</td>
<td>-802.57</td>
<td>0.2593</td>
</tr>
<tr>
<td>Average rating of top 4 most helpful reviews</td>
<td>NA</td>
<td>NA</td>
<td>1517.23</td>
<td>0.0099 **</td>
</tr>
<tr>
<td>Average rating of top 10 most helpful reviews</td>
<td>NA</td>
<td>NA</td>
<td>-9544.52</td>
<td>0.0033 **</td>
</tr>
<tr>
<td>Average rating for top 10 most recent reviews</td>
<td>NA</td>
<td>NA</td>
<td>8511.05</td>
<td>0.0076 **</td>
</tr>
</tbody>
</table>

**Content Analysis**

Figure 4 shows the numbers of active app developers in each group over the four time periods. There were more app developers in the Apple market than in the Google market. Similar to the trend of app increase rate shown in Figure 1, the app developers grew steady in the Apple market while the developers grew two to three times faster in the Google market. In both markets, an app developer created two apps on average. Moreover, more than half of the apps were created by a small group of developers who had two or more apps. For example, for the Health and Fitness apps on the Apple App Store, 53% of the apps were created by 11% of the app developers with two or more apps, while 47% of the apps were created by 89% of the app developers with only one app.

**Figure 4. Distribution of active app developers**
The content analysis continued to investigate potential copy and paste behaviors since many apps were created by a small group of developers. The copy-paste behaviors were confirmed by the high (more than 50%) text similarity of the app descriptions for developers who created at least two apps in both markets. In addition, the analysis of app descriptions changes from time to time indicated that ME apps had a large variance of changes in both markets. Furthermore, only 13.86% of apps had a description change that led to a lower RRL. In other words, most description changes resulted in the same or even a higher RRL.

After reviewing the selected 100 apps which had a readability level of eighth grade or above (higher than the recommended level), two common factors contributing to the low readability of app descriptions were identified. 1) Lack of Writing Standard. Unlike online health education materials, there is no standard way to write app descriptions on either market. Therefore, the writing styles had high variations and app descriptions may contain repetitive content and unnecessary information. For example, an app description may contain a ‘To-Do list’ (Table 6 left, App 1) with different subjects or the change history of the app. In these cases, many difficult words (words with many syllables) were included, thus increasing the complexity of reading. App descriptions can be revised in a more succinct and organized way to improve the readability, which may favor the readability measure in improving the sentence length and the syllable count. 2) Search Optimization. Similar to many existing web search engines, the app stores have their own mechanism to retrieve relevant apps for users given a query. The search mechanism may consider app title, description, user ratings, and other metadata, and help users uncover the apps they need. Therefore, there are some App Store Optimization approaches\textsuperscript{32} to boost the search results in the app stores. Obviously, some app developers would employ these strategies and incorporated many popular keywords (Table 6 right, App 2) in the descriptions, and expected that these approaches will bring more awareness and traffic to their apps. However, these approaches may also make the app descriptions less readable, or negatively impact the readability of app descriptions.

Table 6. App description examples.

<table>
<thead>
<tr>
<th>Partial description of App 1*</th>
<th>Partial description of App 2*</th>
</tr>
</thead>
<tbody>
<tr>
<td>48) How to Do Bojutsu Kamae Basic Postures</td>
<td>Mexican Fideo Soup</td>
</tr>
<tr>
<td>49) How to Do the Bojutsu Striking Drill</td>
<td>Creamy Asparagus Soup</td>
</tr>
<tr>
<td>50) How to Do the Upper Level Bojutsu Block</td>
<td>Chicken and Saffron Rice Soup</td>
</tr>
<tr>
<td>51) How to Do Goho from Bojutsu Training</td>
<td>Chicken, Wild Rice, and Mushroom Soup</td>
</tr>
<tr>
<td>52) Bo Furi Gata from Bojutsu Training</td>
<td>Mustard Greens Soup</td>
</tr>
<tr>
<td>53) How to Do the Muto Dori Technique</td>
<td>Soup-Ojai Valley Inn Tortilla Soup</td>
</tr>
<tr>
<td>54) How to Do the Kenjutsu Technique</td>
<td>Great collection of cheesecake recipe app,banana bread,chicken breast recipes,pumpkin pie recipe,dessert recipes,soup recipes,dinner recipes,easy chicken recipes,sweet potato pie,apple crisp recipe,apple butter recipe,banana cake,italian recipes,peach pie,cookie recipes app,waffle recipe app,dessert recipes app,cupcake recipes app,Baking recipes app,Chinese Recipes app,Chocolate recipes app,Delicious recipes app,chili recipe,Dessert Recipes app,Dinner recipe app,Fish recipes app,Grill Recipes app,Indian recipes… app,Italian Recipes app,Kids recipes app…</td>
</tr>
<tr>
<td>55) How to Do Kusari-Fundo Techniques</td>
<td></td>
</tr>
<tr>
<td>56) How to Use Hanbo Techniques</td>
<td></td>
</tr>
<tr>
<td>57) How to Use Metsubushi Techniques</td>
<td></td>
</tr>
<tr>
<td>58) How to Throw a Shuriken</td>
<td></td>
</tr>
<tr>
<td>59) How to Throw the Bo Shuriken</td>
<td></td>
</tr>
<tr>
<td>60) How to Use the Naginata</td>
<td></td>
</tr>
<tr>
<td>61) How to Use the Yari</td>
<td></td>
</tr>
<tr>
<td>62) How to Use the Kusarigama</td>
<td></td>
</tr>
<tr>
<td>63) How to Use a Rope</td>
<td></td>
</tr>
</tbody>
</table>

*The app descriptions were extracted from the mHealth app repository, not the most current information. Both apps were on the Google Play Store.

Discussion

In this study, we characterized the mHealth apps in both Apple App Store and Google Play Store and summarized the readability of app descriptions. Although there were more apps on Apple App Stores, the Google Play Store was fast growing, especially in the Health and Fitness category. The overall RRL of app descriptions was between eleventh- and fifteenth-grade level, which was far beyond the recommended third- to seventh-grade level. This
finding was consistent with a recent study indicating that the privacy policies of mHealth apps were not very readable (sixteenth-grade level)\(^3\). Moreover, only around 6\% apps that meet the recommended grade level highlights the need for improvement.

We also explored price and content rating and their relationships with description readability. Interestingly, paid apps did not provide better readability than free apps, and the content ratings of apps did not always match the corresponding grade levels. In other words, apps with suitable content may not be as readable to the target users as they are supposed to be and unfortunately become unsuitable. There could be other predictors. For example, expert involvement may increase the app installs\(^34\). Another key factor in the choice of apps may be the ability of a mHealth app to manage patient-generated health data given the current advances and wide adoption of smart watches and wearable devices. Developers and vendors should consider these factors carefully in app development and advertisement.

The content analysis of the less readable apps revealed that a small group of app developers created two or more apps and contributed to a significant portion of the apps. These app developers may copy and paste their app descriptions for efficiency and they have no incentive to improve the readability of app descriptions over time. The high RRL of app descriptions may be contributed by the lack of writing standard and the desire for search optimization, which is a similar pattern identified in a previous study analyzing cardiological apps in the Apple App Store in German\(^35\). More research is needed to demonstrate the best practice of writing app descriptions, balancing between readability, retrievability, and efficiency.

This study had several strengths. Based on our knowledge, this is the first study conducting both retrospective and prospective analyses with a focus on the readability of mHealth app descriptions in both Apple and Google market. Moreover, this study adopted multiple methods. Not only did it conduct statistical summary and group comparisons, but also it developed a model to predict user behaviors as well as reviewed app descriptions to understand potential causes for low readability. This study also had a few limitations. First, we only analyzed ‘Health and Fitness’ and ‘Medical’ categories based on the data obtained from the previous study\(^9\), acknowledging that there could be other related (or new) categories that we missed. Second, different app markets have different developer groups, description writing guidelines, and management policies, which affect our analysis results. Moreover, since we only focused on Apple Apple Store and Google Play Store, and the results may not be generalizable to other app markets (e.g. Amazon Kindle Store). Third, the study did not consider that fact that users install apps but delete them afterwards due to the lack of usefulness. Also, the study did not conduct a thorough content analysis to understand how the readability of app descriptions may impact the end users. Next, the modeling of app installs and their metadata only showed weak to moderate effect size and can be improved. In addition, the classic readability measures used in the present study have known limitations. For example, a word with multiple syllables may not be difficult and some short words can be very difficult to read and understand. Last but not least, the readability of app descriptions was not validated by app users. Future research can conduct user studies to validate the readability scores as well as understand user’s perspectives on less readable app descriptions.

**Conclusion**

We assessed the readability of mHealth app descriptions and demonstrated its potential role in mHealth app choices. This study serves as the first attempt to address this issue and enabled future studies to develop solutions to improve the readability of app description and further help lay persons select suitable mHealth apps for their health needs.

**Acknowledgements**

We thank Dr. Welong Xu and Dr. Yin Liu for their willingness to share the mHealth app repository. We also thank Mr. Karthikeyan Meganathan at the University of Cincinnati College of Medicine for his assistance on the statistical analysis and Ms. Anunita Nattam for her efforts in proofreading the manuscript.

**References**


The Addition of United States Census-TRACT Data Does Not Improve the Prediction of Substance Misuse

Daniel To, BA\textsuperscript{1}, Cara Joyce, PhD\textsuperscript{2}, Sujay Kulshrestha, MD\textsuperscript{3}, Brihat Sharma, MS\textsuperscript{4}, Dmitry Dligach, PhD\textsuperscript{2,5}, Matthew Churpek, MD, MPH, PhD\textsuperscript{6}, Majid Afshar, MD, MSCR\textsuperscript{6}

\textsuperscript{1}Stritch School of Medicine, Loyola University Chicago, Maywood, IL; \textsuperscript{2}Department of Public Health, Stritch School of Medicine, Loyola University Chicago, Maywood, IL; \textsuperscript{3}Department of Surgery, Loyola University Medical Center, Maywood, IL; \textsuperscript{4}Department of Psychiatry and Behavioral Sciences, Rush University Medical Center, Chicago, IL; \textsuperscript{5}Department of Computer Science, Loyola University Chicago, Chicago, IL; \textsuperscript{6}Department of Medicine, University of Wisconsin School of Medicine and Public Health, Madison, Wisconsin

ABSTRACT

Predictors from the structured data in the electronic health record (EHR) have previously been used for case-identification in substance misuse. We aim to examine the added benefit from census-tract data, a proxy for socioeconomic status, to improve identification. A cohort of 186,611 hospitalizations was derived between 2007 and 2017. Reference labels included alcohol misuse only, opioid misuse only, and both alcohol and opioid misuse. Baseline models were created using 24 EHR variables, and enhanced models were created with the addition of 48 census-tract variables from the United States American Community Survey. The absolute net reclassification index (NRI) was applied to measure the benefit in adding census-tract variables to baseline models. The baseline models already had good calibration and discrimination. Adding census-tract variables provided negligible improvement to sensitivity and specificity and NRI was less than 1\% across substance groups. Our results show the census-tract added minimal value to prediction models.

INTRODUCTION

Substance misuse is a common cause of hospitalization and death in the United States. The most common type of primary and secondary substance-related diagnosis among inpatient hospitalizations is alcohol-related disorders, and it ranks second in 7-day readmission rates\textsuperscript{1}. Additionally, rates of opioid-related deaths have continued to increase, particularly with the advent of synthetic opioids. Approximately 70\% of drug overdose deaths in 2018 involved opioids\textsuperscript{2}. Both alcohol and opioid misuse are complex behavioral conditions that encompass a variety of co-existing conditions and social determinants of health.

Both alcohol and opioid misuse have been shown to be associated with social and behavioral determinants of health, such as poverty level\textsuperscript{3}, education level\textsuperscript{4}, and employment status\textsuperscript{5}. Furthermore, substance use outcomes appear to be clustered by geographic area\textsuperscript{6}. One community study using geographical information software suggests its benefit in characterizing drug use in neighborhoods\textsuperscript{7}. This suggests that environmental influences may play a role in substance use. Measures of census-level socioeconomic status (SES) indicators function as proxies for individual-level SES information, help fill a gap in electronic health record (EHR) data\textsuperscript{8}, and improve the accuracy for identifying cases of substance misuse in patients. The SES indicators provide additional value beyond individual risk factors in predicting health risk and examining health outcomes\textsuperscript{9,10}. Few studies have linked EHR data with census-level data for substance misuse, so their effectiveness is less apparent for prediction.

In this study, we used the publicly available United States American Community Survey data summarized to census-tract to represent SES for patients. We linked the census-tract variables to geo-coded patient addresses in the EHR at a tertiary care health system to examine the added benefit in the census-tract data to existing structured EHR data. We aim to study the added benefit of the census-tract data for the prediction of alcohol misuse, opioid misuse, and both alcohol and opioid misuse. We hypothesize that census-tract data will improve the net reclassification of cases for each type of substance misuse over EHR data alone.
METHODS

Patient Setting
Loyola University Medical Center (LUMC) is a 559-bed hospital and tertiary academic center, including a burn and Level 1 trauma center serving Chicago and its western suburbs. LUMC has maintained Epic (Epic Systems Corporation, Verona, Wisconsin) as its EHR vendor since 2003 and includes a Microsoft SQL server-based clinical data warehouse (CDW) that has been available for research since 2007. The study was performed at an encounter level. The study population is composed of all adult (≥18 years of age) inpatient encounters between January 1, 2007 and September 30, 2017. Exclusion criteria were the following: (1) outpatient encounters and (2) encounters where census-tract data could not be matched to patient address.

Reference labels for substance misuse
Misuse included patients with opioid use disorder, taking an illicit opioid or non-prescribed opioid, alcohol use disorder, and excessive alcohol consumption as defined by National Institute on Alcohol Abuse and Alcoholism. Two methods were used to identify cases. First, a combination of International Classification of Diseases (ICD)-9 and –10 codes for opioid misuse and alcohol misuse were adopted from the Healthcare Cost and Utilization Project (HCUP) ICD codes for opioid abuse, opioid dependence, opioid poisoning, alcohol dependence, and alcohol abuse. A total of 20 ICD codes were used for opioid misuse and 22 ICD codes for alcohol misuse. Second, computable phenotypes that used natural language processing were applied to the clinical notes of the EHR for both alcohol and opioid misuse. The computable phenotypes for opioid misuse and alcohol misuse had previously been trained and validated at LUMC and they both had an area under the receiver operating characteristic curve (AUROC) of greater than 0.9011-13. All the data (ICD codes and clinical notes) used to build the reference labels were independent from the variables that were used as features in the models below.

Candidate Variables from the EHR
Variables for analyses from the EHR were extracted from the following domains: (1) demographics including insurance status; (2) comorbidities organized by Elixhauser disease classification categories and present on admission (excluding the codes used in the reference labels)14; (3) Elixhauser readmission score and Elixhauser mortality score; (4) inpatient pain score from the admission nursing flowsheets; (5) ICD codes for chronic pain present on admission; (6) laboratory testing with blood alcohol concentration (BAC) in mg/dL. A total of 24 EHR variables were examined. Median imputation for integer variables and mode imputation for nominal variables were applied to integer values with missing data except for BAC which was categorized as not tested.

Candidate Variables from Census Tract Data
The United States (US) Census Tract socioeconomic (SES) data were used as a proxy for individual-level social and behavioral determinants of health. The ‘censusapi’ R package was used as a wrapper for the US Census Bureau’s Application Program Interfaces (API)15. The Census APIs were used to match the addresses to corresponding geocodes for all patients in our analytic cohort16. The data were extracted from the American Community Survey 5-year Data between 2013 and 201717. Forty-eight census-tract variables for analyses were extracted from the following domains: (1) demographics; (2) highest education level; (3) marital status; (4) household composition; (5) insurance status; (6) employment status; (7) first language; (8) veteran status; (9) percent of households below poverty level.

Statistical analysis: Association of EHR and census-tract variables with substance misuse
Individual EHR and census-tract variables were examined across groups by substance use type (alcohol misuse only, opioid misuse only, alcohol and opioid misuse, and no misuse) (Tables 1 and 2). The variables were entered into a generalized linear mixed effects model with Poisson distribution and included random intercepts to account for within-patient correlation due to multiple inpatient encounters over time. A total of 72 candidate variables were examined representing both the structured EHR variables and census-tract variables. Results from all substance use models were reported using prevalence ratios with 95% confidence intervals (CI).

Predictive Analytics
The dataset was split into 70% (n=130,628) for training and 30% (n=55,983) for testing. Variable selection was performed using the least absolute shrinkage and selection operator (LASSO) in a GLM and hyperparameter tuning to find the largest value of $\lambda$ that is within one standard error of the minimum was performed on the training set using 10-fold cross-validation. First, baseline EHR models were derived to select features from the candidate EHR variables for predicting each type of substance misuse (alcohol misuse vs. no misuse; opioid misuse vs. no misuse; both types
vs. no misuse). Second, the census-tract variables were added to the list of candidate variables and variable selection with LASSO was performed. The Area Under the Receiver Operating Characteristics (AUROC) was calculated for each model, and the AUROC between the baseline and enhanced models were compared using a bootstrap test for two correlated ROC curves with 100 permutations. The following formula was used: $D=(\text{baseline AUROC} - \text{enhanced AUROC})/s$ where $s$ is the standard deviation of the bootstrap differences and $D$ is compared to the normal distribution. In addition to examining discrimination with AUROC, we also examined calibration with the calibration slope and intercept with their 95% CIs.

The absolute Net Reclassification Index (NRI) was applied to examine the benefit in adding census-tract variables to EHR data across multiple thresholds on the AUROC. The thresholds examined were the following: (1) Youden’s J index (maximizing accuracy and minimizing error); (2) the highest sensitivity/recall threshold when specificity was set to 75%, and (3) the highest specificity threshold when sensitivity/recall was set to 75%. The highest absolute NRI from these possible thresholds was reported in the results. The absolute NRI represents the absolute number of patients correctly reclassified by the enhanced model over the EHR-only model. It was used to determine if the enhanced model performed better than the EHR-only model. The net reclassification was calculated by subtracting the incorrect reclassification of the enhanced model and the correct reclassification enhanced model. An incorrect reclassification is when the baseline model accurately classified the case, but the enhanced model incorrectly reclassifies the case. Similarly, a correct reclassification is when the baseline incorrectly classified the case, but the nested model correctly reclassifies the case. The absolute NRI can be calculated by the following equation:

$$\text{Absolute NRI} = \frac{\text{misuse net reclassification} + \text{no misuse net reclassification}}{\text{total encounters}} \times 100$$

In addition, classification plots were built to better visualize discrimination and compare AUROCs. Classification plots also overcome the problem of comparing model performance conditional on specific thresholds by showing all true positive (sensitivity) and false positive rates (specificity) by risk thresholds. Figure 1 shows sensitivity/recall and specificity conditional on all risk thresholds.

The analysis was performed using RStudio Version 1.1.463 (RStudio Team, Boston, MA). All results from the prediction models are reported on the test dataset. The Institutional Review Board of Loyola University Chicago approved this study.

RESULTS

Patient Characteristics

The census tract variables for SES could not be linked in 16.2% (n=37,254) of patient encounters due to missing or incorrect address information in the EMR. The final cohort analyzed was composed of 186,611 adult hospitalizations. There were 13,263 (7.1%) positive cases of alcohol misuse, 4,484 (2.4%) positive cases of opioid misuse, and 2,896 (1.6%) cases of concurrent alcohol and opioid misuse. The association of different patient characteristics derived from the EHR data is listed in Table 1. The Elixhauser comorbidity most strongly associated with alcohol misuse was liver disease at 3.26 (95% CI, 3.07-3.46), followed by psychosis at 2.30 (95% CI, 2.14-2.47). Psychosis was the most strongly associated with opioid misuse and combined alcohol-opioid substance use groups with a prevalence ratio of 2.66 (95% CI, 2.40-2.95) and 2.50 (95% CI, 2.13-2.94), respectively. Discharge to a psychiatric facility was strongly associated with all categories of substance use (p<0.01 for all substance use types). Detectable BAC levels of <80 mg/dL and >80 mg/dL (above legal limit) had the strongest association for alcohol misuse with prevalence ratios of 9.25 (95% CI, 6.30-11.29) and 8.43 (95% CI, 8.55-10.01), respectively.

Variables from major domains of the census-tract variables from the 2013-2017 American Community Survey are listed in Table 2. Higher levels of per capita income, median household earnings, and median income were associated with lower prevalence of each substance use type (p<0.01 for all comparisons). Similarly, greater levels of poverty were associated with higher rates of substance misuse (p<0.01 for all comparisons). Further, lower levels of high school education and increases in food stamp usage had a positive association with all substance use types (p<0.01 for all comparisons).
### Table 1. Patient Demographics and Characteristics across groups for Substance Misuse

<table>
<thead>
<tr>
<th></th>
<th>Alcohol Misuse Only (n=13263)</th>
<th>Opioid Misuse Only (n=4484)</th>
<th>Alcohol + Opioid Misuse (n=2896)</th>
<th>No Misuse (n=165968)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Prevalence</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ratio (95% CI)</td>
<td>P Value</td>
<td>Ratio (95% CI)</td>
<td>P Value</td>
</tr>
<tr>
<td><strong>Age (years)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18-24 (referent)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25-34</td>
<td>0.98 (0.87, 1.10)</td>
<td>0.76</td>
<td>1.03 (0.84, 1.28)</td>
<td>0.75</td>
</tr>
<tr>
<td>35-44</td>
<td>1.05 (0.94, 1.18)</td>
<td>0.41</td>
<td>1.13 (0.91, 1.40)</td>
<td>0.28</td>
</tr>
<tr>
<td>45-54</td>
<td>1.15 (1.03, 1.29)</td>
<td>0.01</td>
<td>0.97 (0.78, 1.20)</td>
<td>0.78</td>
</tr>
<tr>
<td>55+</td>
<td>0.54 (0.48, 0.59)</td>
<td>&lt;0.001</td>
<td>0.52 (0.42, 0.63)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td><strong>Sex</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>3.09 (2.85, 3.31)</td>
<td>&lt;0.001</td>
<td>1.74 (1.55, 1.95)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td><strong>Race/Ethnicity</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-Hispanic White</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hispanic</td>
<td>1.17 (1.03, 1.32)</td>
<td>0.01</td>
<td>0.97 (0.73, 1.28)</td>
<td>0.83</td>
</tr>
<tr>
<td>Other</td>
<td>1.14 (1.05, 1.24)</td>
<td>&lt;0.001</td>
<td>0.85 (0.70, 1.03)</td>
<td>0.10</td>
</tr>
<tr>
<td><strong>Insurance</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Private</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medicare</td>
<td>0.69 (0.65, 0.73)</td>
<td>&lt;0.001</td>
<td>0.80 (0.70, 0.90)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Medicaid</td>
<td>1.97 (1.85, 2.10)</td>
<td>&lt;0.001</td>
<td>2.22 (2.02, 2.43)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>No Insurance/Other</td>
<td>2.88 (2.68, 3.09)</td>
<td>&lt;0.001</td>
<td>2.55 (2.37, 2.85)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td><strong>Comorbidities</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CHF</td>
<td>0.65 (0.61, 0.70)</td>
<td>&lt;0.001</td>
<td>0.95 (0.85, 1.06)</td>
<td>0.36</td>
</tr>
<tr>
<td>Hypertension</td>
<td>0.65 (0.62, 0.68)</td>
<td>&lt;0.001</td>
<td>0.83 (0.76, 0.90)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Neurological</td>
<td>1.46 (1.39, 1.54)</td>
<td>&lt;0.001</td>
<td>1.32 (1.21, 1.44)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Pulmonary</td>
<td>1.03 (0.97, 1.09)</td>
<td>0.39</td>
<td>1.19 (1.07, 1.31)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Complicated DM</td>
<td>0.70 (0.64, 0.75)</td>
<td>&lt;0.001</td>
<td>0.92 (0.81, 1.05)</td>
<td>0.22</td>
</tr>
<tr>
<td>Uncomplicated DM</td>
<td>0.74 (0.69, 0.79)</td>
<td>&lt;0.001</td>
<td>0.83 (0.75, 0.92)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Renal</td>
<td>0.68 (0.68, 0.78)</td>
<td>0.001</td>
<td>0.84 (0.74, 0.93)</td>
<td>0.002</td>
</tr>
<tr>
<td>Liver</td>
<td>3.26 (3.07, 3.46)</td>
<td>&lt;0.001</td>
<td>1.97 (1.85, 2.14)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>HIV</td>
<td>1.27 (0.94, 1.72)</td>
<td>0.12</td>
<td>1.80 (1.20, 2.69)</td>
<td>0.005</td>
</tr>
<tr>
<td>Rheumatic</td>
<td>0.45 (0.37, 0.54)</td>
<td>&lt;0.001</td>
<td>0.80 (0.60, 1.06)</td>
<td>0.12</td>
</tr>
<tr>
<td>Obesity</td>
<td>0.62 (0.58, 0.67)</td>
<td>&lt;0.001</td>
<td>0.99 (0.88, 1.11)</td>
<td>0.83</td>
</tr>
<tr>
<td>Weight Loss</td>
<td>1.30 (1.22, 1.40)</td>
<td>&lt;0.001</td>
<td>1.19 (1.06, 1.33)</td>
<td>0.002</td>
</tr>
<tr>
<td>Anemia</td>
<td>0.98 (0.93, 1.03)</td>
<td>0.40</td>
<td>1.05 (0.98, 1.13)</td>
<td>0.19</td>
</tr>
<tr>
<td>Psychosis</td>
<td>2.30 (2.14, 2.47)</td>
<td>&lt;0.001</td>
<td>2.66 (2.40, 2.95)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Depression</td>
<td>1.41 (1.33, 1.49)</td>
<td>&lt;0.001</td>
<td>1.82 (1.67, 1.97)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td><strong>Disposition</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acute Care</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>In-Hospital Death</td>
<td>1.26 (1.12, 1.42)</td>
<td>&lt;0.001</td>
<td>0.69 (0.54, 0.86)</td>
<td>0.001</td>
</tr>
<tr>
<td>Home</td>
<td>0.92 (0.86, 0.98)</td>
<td>0.01</td>
<td>0.93 (0.83, 1.04)</td>
<td>0.18</td>
</tr>
<tr>
<td>AMA</td>
<td>1.68 (1.45, 1.95)</td>
<td>&lt;0.001</td>
<td>1.31 (1.09, 1.58)</td>
<td>0.004</td>
</tr>
<tr>
<td>Long Term Care</td>
<td>1.02 (0.93, 1.12)</td>
<td>0.67</td>
<td>0.93 (0.79, 1.09)</td>
<td>0.36</td>
</tr>
<tr>
<td>Psychiatric Hospital</td>
<td>4.81 (4.08, 5.67)</td>
<td>&lt;0.001</td>
<td>3.08 (2.40, 3.95)</td>
<td>0.70</td>
</tr>
<tr>
<td><strong>Alcohol Testing</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not Tested</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BAC = 0</td>
<td>3.48 (3.28, 3.68)</td>
<td>&lt;0.001</td>
<td>2.30 (2.08, 2.54)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Below Legal (≤80)</td>
<td>9.25 (6.30, 11.29)</td>
<td>&lt;0.001</td>
<td>3.36 (2.85, 3.95)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Above Legal (&gt;80)</td>
<td>8.43 (8.55, 10.01)</td>
<td>&lt;0.001</td>
<td>3.15 (2.78, 3.56)</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

CHF = Congestive Heart Failure; DM = Diabetes; AMA = Left against medical advice; HIV = Human Immunodeficiency Virus; BAC = Blood alcohol level; Acute care = another short-term general hospital for inpatient care, home health service, immediate care facility; Chronic care = inpatient rehab facility, nursing facility, long-term care hospital, skilled nursing facility; Chronic pain not part of Elixhauser codes – the ICD-9/10 code used: 338, 338.0, 338.2, 338.21, 338.22, 338.28, 338.29, 338.4, 724.5, G90, G90.0, G90.2, G90.21, G90.22, G90.23, G90.29, G90.3, G90.4, REZ; Alcohol level was shown to represent an important indicator of census-level SES that correlates well with other SES measures therefore, we categorized patients into low (<9.9 percent of households below federal poverty level), middle (10.0–19.9 percent of households below federal poverty level), or high-poverty census-tract (20.0+ percent of households below federal poverty level).
Table 2. Census tract data from the 2013-2017 American Community Survey linked to hospitalizations from the electronic health record

<table>
<thead>
<tr>
<th>Socioeconomic status</th>
<th>Alcohol Misuse Only (n=13263)</th>
<th>P Value</th>
<th>Prevalence Ratio (95% CI)</th>
<th>P Value</th>
<th>Prevalence Ratio (95% CI)</th>
<th>P Value</th>
<th>Prevalence Ratio (95% CI)</th>
<th>P Value</th>
<th>Prevalence Ratio (95% CI)</th>
<th>P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less than High School Education</td>
<td>1.01 (1.01, 1.02)</td>
<td>&lt;0.001</td>
<td>1.01 (1.00, 1.01)</td>
<td>0.03</td>
<td>1.02 (1.01, 1.03)</td>
<td>&lt;0.001</td>
<td>1.00 (1.00, 1.00)</td>
<td>&lt;0.001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Marriage Status</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Married</td>
<td>0.99 (0.98, 0.99)</td>
<td>&lt;0.001</td>
<td>0.98 (0.96, 0.99)</td>
<td>&lt;0.001</td>
<td>0.97 (0.96, 0.98)</td>
<td>&lt;0.001</td>
<td>1.00 (1.00, 1.00)</td>
<td>&lt;0.001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Never Married</td>
<td>1.01 (1.01, 1.02)</td>
<td>&lt;0.001</td>
<td>1.02 (1.01, 1.03)</td>
<td>&lt;0.001</td>
<td>1.03 (1.02, 1.04)</td>
<td>&lt;0.001</td>
<td>1.00 (1.00, 1.00)</td>
<td>&lt;0.001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Food Stamp Usage (10% increase)</td>
<td>1.15 (1.12, 1.18)</td>
<td>&lt;0.001</td>
<td>1.21 (1.14, 1.28)</td>
<td>&lt;0.001</td>
<td>1.32 (1.22, 1.42)</td>
<td>&lt;0.001</td>
<td>0.98 (0.97, 0.98)</td>
<td>&lt;0.001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Disability (5% increase)</td>
<td>1.02 (1.01, 1.03)</td>
<td>&lt;0.001</td>
<td>1.03 (1.02, 1.05)</td>
<td>0.001</td>
<td>1.04 (1.02, 1.07)</td>
<td>0.001</td>
<td>1.00 (1.00, 1.00)</td>
<td>&lt;0.001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Homeowner (5% increase)</td>
<td>0.96 (0.95, 0.97)</td>
<td>&lt;0.001</td>
<td>0.95 (0.94, 0.72)</td>
<td>&lt;0.001</td>
<td>0.93 (0.90, 0.95)</td>
<td>&lt;0.001</td>
<td>1.01 (1.00, 1.01)</td>
<td>&lt;0.001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not a Citizen of US (5% increase)</td>
<td>1.06 (1.04, 1.08)</td>
<td>&lt;0.001</td>
<td>0.97 (0.92, 1.02)</td>
<td>0.25</td>
<td>1.04 (1.04, 1.04)</td>
<td>&lt;0.001</td>
<td>0.99 (0.99, 1.00)</td>
<td>&lt;0.001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Per Capita Income (per $10,000)</td>
<td>0.87 (0.85, 0.89)</td>
<td>&lt;0.001</td>
<td>0.87 (0.82, 0.92)</td>
<td>&lt;0.001</td>
<td>0.78 (0.71, 0.85)</td>
<td>&lt;0.001</td>
<td>1.02 (1.01, 1.02)</td>
<td>&lt;0.001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Median Earnings (per $10,000)</td>
<td>0.87 (0.85, 0.89)</td>
<td>&lt;0.001</td>
<td>0.87 (0.82, 0.92)</td>
<td>&lt;0.001</td>
<td>0.76 (0.70, 0.84)</td>
<td>&lt;0.001</td>
<td>1.02 (1.01, 1.02)</td>
<td>&lt;0.001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Median Household Income (per $10,000)</td>
<td>0.95 (0.94, 0.96)</td>
<td>&lt;0.001</td>
<td>0.94 (0.92, 0.97)</td>
<td>&lt;0.001</td>
<td>0.90 (0.86, 0.94)</td>
<td>&lt;0.001</td>
<td>1.01 (1.01, 1.01)</td>
<td>&lt;0.001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Poverty Level</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low (≤ 9.9%)</td>
<td>(referent)</td>
<td></td>
<td>(referent)</td>
<td></td>
<td>(referent)</td>
<td></td>
<td>(referent)</td>
<td></td>
<td>(referent)</td>
<td></td>
</tr>
<tr>
<td>Middle (10%-19.9%)</td>
<td>1.22 (1.13, 1.32)</td>
<td>&lt;0.001</td>
<td>1.19 (1.01, 1.41)</td>
<td>0.04</td>
<td>1.47 (1.15, 1.87)</td>
<td>0.002</td>
<td>0.97 (0.96, 0.98)</td>
<td>&lt;0.001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High (≥ 20%)</td>
<td>1.53 (1.39, 1.68)</td>
<td>&lt;0.001</td>
<td>1.68 (1.38, 2.05)</td>
<td>&lt;0.001</td>
<td>2.03 (1.53, 2.70)</td>
<td>&lt;0.001</td>
<td>0.95 (0.93, 0.96)</td>
<td>&lt;0.001</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Alcohol-Only Model

The EHR variables selected by LASSO to create the baseline model were BAC, sex, insurance status, all the Elixhauser comorbidities, pain level, and Elixhauser readmission and mortality indices. The 23-variable model had an AUROC of 0.879 (95% CI, 0.874-0.885). The calibration slope and intercept were 1.11 (95% CI, 1.08-1.13) and 0.22 (95% CI, 0.17-0.28), respectively.

The additional SES variables selected by LASSO for the enhanced model for alcohol misuse were the following from the patient’s census-tract: (1) proportion divorced; (2) proportion veterans; (3) proportion without a high school degree; (4) proportion college graduate; (5) proportion with household size of two; (6) proportion homeowner; (7) proportion not in labor force; and (8) proportion 25-64 years old. Only one Elixhauser comorbidity was removed from the baseline model. The 32-variable model had a small improvement with an AUROC of 0.880 (95% CI: 0.875-0.886), and a p-value <0.01 for comparison between the baseline and nested models. A similar calibration slope and intercept were found at 1.10 (95% CI, 1.08-1.13) and 0.21 (95% CI, 0.15-0.27), respectively. The threshold that provided the maximal benefit for reclassification had an absolute NRI of 0.39% (Table 3). Figure 1a represents the classification plots and shows negligible gains across risk thresholds for sensitivity and specificity.

Opioid-Only Model

The EHR variables selected by LASSO to create the baseline model were BAC, age, race and ethnicity, sex, insurance status, congestive heart failure, neurological disorders, pulmonary disorders, uncomplicated diabetes, complicated diabetes, renal disorders, liver disorders, HIV, metastasis, tumor, rheumatic disorders, obesity, weight loss, anemia, psychosis, depression, and Elixhauser readmission and mortality indices. The 23-variable model had an AUROC of 0.857 (95% CI, 0.847-0.866). The calibration slope and intercept were 1.14 (95% CI, 1.10-1.17) and 0.42 (95% CI, 0.33-0.52), respectively.
The additional SES variables selected by LASSO for the enhanced model for opioid misuse were the following from the patient’s census tract: (1) proportion black; (2) proportion white, (3) per capita income; (4) proportion food stamps, and (5) median earnings. For the enhanced opioid misuse model, the 28-variable model had an AUROC of 0.857 (95% CI, 0.848-0.866) and no improvement in the AUROC was found over the baseline model (p=0.65). The enhanced opioid misuse model and relatively no change in the calibration slope and intercept of 1.14 (95% CI, 1.1-1.17) and 0.44 (95% CI, 0.35-0.54), respectively. The threshold that provided the maximal benefit in reclassification had an absolute NRI of 0.04% (Table 3). Figure 1b represents the classification plots with no appreciable change visualized across risk thresholds for sensitivity and specificity.

**Alcohol and Opioid Model**

The EHR variables selected by LASSO to create the baseline model were BAC, age, sex, all the Elixhauser comorbidities, pain level, and Elixhauser readmission and mortality indices. For the baseline alcohol and opioid misuse model, a 23-variable model was derived with an AUROC of 0.952 (95% CI, 0.945-0.960). The baseline model had a calibration slope and intercept of 1.14 (95% CI, 1.10-1.19) and 0.54 (95% CI, 0.40-0.67), respectively.

The additional SES variables selected by LASSO to create the enhanced model were the following from the patient’s census tract: (1) proportion married; (2) proportion white; (3) proportion household size of two; (4) proportion disabled; (4) median earning. For the enhanced alcohol and opioid misuse model, a 29-variable model was derived with an AUROC of 0.953 (0.946-0.960) and no improvement in the AUROC was found over the baseline model (p=0.21). The model had minimal change in the calibration slope and intercept at 1.14 (95% CI, 1.10-1.19) and 0.53 (95% CI, 0.40-0.66), respectively. None of the thresholds examined provided any benefit in reclassification with the best absolute NRI at -0.13% (Table 3). Figure 1c represents the classification plots with no improvements visualized across risk thresholds for sensitivity and specificity.

**Table 3. Net Reclassification after the addition of selected census-tract variables**

<table>
<thead>
<tr>
<th></th>
<th>No Alcohol Misuse (n=51069)</th>
<th>Alcohol Misuse (n=4914)</th>
<th>No Opioid Misuse (n=53662)</th>
<th>Opioid Misuse (n=2321)</th>
<th>No Alcohol and Opioid Misuse (n=55028)</th>
<th>Alcohol and Opioid Misuse (n=955)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct Reclassification:</td>
<td>304</td>
<td>12</td>
<td>605</td>
<td>21</td>
<td>26</td>
<td>14</td>
</tr>
<tr>
<td>Incorrect Reclassification:</td>
<td>39</td>
<td>58</td>
<td>558</td>
<td>45</td>
<td>107</td>
<td>7</td>
</tr>
<tr>
<td>Net Reclassification</td>
<td>265</td>
<td>-46</td>
<td>47</td>
<td>-24</td>
<td>-81</td>
<td>7</td>
</tr>
</tbody>
</table>

All results reported for the test dataset (n=55,983).
DISCUSSION

Prior studies have shown that using readily available data in the EHR may be useful in the identification of individuals with substance misuse. Because substance misuse is strongly associated with SES, we added census-tract variables, a proxy for SES data, to the EHR-only model. Our results show that including census-tract variables into the prediction model using a LASSO approach resulted in several census-tract variables being added to the model but only with small gains in AUROC for the alcohol misuse model. For the most part, there were little to no gains to the AUROC, absolute NRI, and across risk thresholds for sensitivity/recall and specificity. Our results indicate that the enhanced model did not contribute much predictive value, but our models from EHR-only data already had baseline AUROCs above 0.84 so there may have been limited capacity for performance gains from baseline. Overall, our models had good discrimination and calibration, but we show little benefit in the added complexity of linking SES data for computable phenotypes in substance misuse. Other computable phenotypes with risk factors in SES may still benefit from the addition of census-tract variables but they are likely on a case-by-case basis.
Our univariable analysis demonstrates that patient’s lower SES status is strongly associated with substance misuse, especially across census-tract data for income, employment, education level, and housing. These data are consistent with individual-level data and consistent with the strong association for substance misuse in patients with Medicaid and uninsured status. From the 72 patient-level and census-tract variables and structured data variables available to our health system, we derived models to predict each substance use type. In addition to the commonly described risk factors in demographics and comorbidities, additional factors such as comorbidities and arriving with a detectable BAC were common to all models. We also found similar characteristics in the patients identified with substance misuse to other urban cohort studies, supporting the reliability in our choice of predictors. Co-substance use, hepatitis, HIV, chronic pain, and mental health conditions are commonly reported risk factors and predictors of substance misuse. For the alcohol misuse model, the Elixhauser codes for neurologic disease include delirium and encephalopathy which are commonly encountered in patients with acute intoxication.

Measures of census-level SES indicators function as proxies for individual-level socioeconomic information and help fill a gap in EHR data. The SES indicators provide additional value beyond individual factors in predicting health risk and examining health outcomes. Few studies have linked EHR data with census-level data for substance misuse so their effectiveness is less apparent for health analyses. The selection operator in our LASSO model did pick approximately a half dozen of the census-tract variables to predict the different types of substance misuse. Across the models, the variables reflected race/ethnicity of the neighborhood, earnings and income status, and disabilities. Health systems are increasingly more accountable for the health of the communities they serve, so these additional data sources may better inform strategies for community outreach and care.

In the end, our baseline models were already well calibrated and started with high performance for discrimination with AUROCs above 0.84. This may explain why our absolute NRI metrics and classification plots across multiple thresholds did not show improvement with the addition of the census-tract variables. This may be viewed as a limitation to our study and should be further explored across other computable phenotypes that are affected by SES. The utility of census-tract variables is focused on patients with substance misuse, so their value may differ in applications to other prediction models. Prior studies have focused on the additive NRI instead of absolute NRI, but it does not consider the prevalence of the cases and non-cases in the cohort and may be misleading. Our study used absolute NRI to account for the low prevalence of cases and represent the total proportion reclassified correctly. The absolute NRI of <1% for all the models indicates little value gained in reclassifications for the added complexity of linking EHR data to the American Community Survey data.

There are several other limitations in our study. First, we assumed that there is little variability in census data between years. The census data used for the modeling were the 5-year average between 2013 and 2017; however, our patient cohort included patients between 2007 and 2017. Additionally, we used the patients’ last known address which may not be representative of the patients’ geographic location during the time of the hospitalization. We did not have accurate addresses on approximately 15% of patients which also included patients experiencing homelessness, which is a major predictor for substance misuse. Because the census-tract variables were derived from the patients’ addresses, the neighborhood characteristics serve only as a proxy for the individual-level characteristics. The result of this study may suggest that the patient’s neighborhood improves the model rather than the patient’s SES. Lastly, this was a single-center study, and an external validation study is needed to determine the generalizability of our results and other model architectures may prove useful.

CONCLUSION

Substance misuse is a behavioral condition that has been shown to be highly associated with SES. However, in this study, we showed that leveraging the publicly available census-tract data, a proxy for SES data, does not improve the substance misuse prediction models. Our results suggest that the census-tract data does not add significant value to our substance misuse computable phenotypes but more work is needed to examine their value across other EHR-level prediction models.

REFERENCES

Increased Clinician Time Using Electronic Health Records During COVID-19 Pandemic

Timothy Tsai, DO 1,2*, Mina Boazak, MD MMCi 1,3*, Eugenia R McPeek Hinz, MD MS 1

*Co-first Authors

Duke University School of Medicine, Durham, NC, USA 1Department of Medicine, 2 Duke Primary Care, 3Department of Psychiatry
Abstract

The COVID-19 pandemic challenged how healthcare systems provided care in socially distanced formats. We hypothesized that the COVID-19 era changes in clinical care delivery models contributed to increased Electronic Health Record (EHR) related work. To evaluate the changes in time and volume metrics of EHR usage, we segregated EHR audit log metric data into PreCOVID2019 March/April/May, initial COVID2020 March/April/May, and late COVID2021 March/April/May for 1262 physician providers. We discovered significant and pragmatically meaningful increases in total average time providers spent in the EHR in minutes mean(SD) PreCOVID2019=1958(1576), MidCOVID2020=1709(1473), LateCOVID2021=2007(1563). Differences in total time in the EHR were significant Pre-mid:p-value=<0.001, but not Pre-Late:p=0.439. Total number of messages received across all specialties increased significantly mean(SD) PreCOVID=459(389), MidCOVID=400(362), LateCOVID 521(423) Pre-Mid p-value=<0.001 and Pre-Late p-value=<0.001. We additionally found changes in total time to differ significantly across select specialties. Based on these findings we recommend further assessment of physician workload and how new factors such as telehealth are contributing to EHR usage.

Introduction

The COVID-19 pandemic has stressed the US healthcare system from increased hospital capacity to new rapidly evolving formats for clinical care delivery. To adapt to these changes, healthcare systems have switched quickly from primarily in-person visits to telemedicine visits(1). The Centers for Medicare & Medicaid Services (CMS) led the way in payment for telehealth services, which led to widespread adoption of telehealth services by other payers (2). The rapid transition from predominantly in-person to telehealth services has led to reports of increases in EHR usage time for clinicians. Moving forward, as healthcare systems are adopting these telehealth practices, it will be important to obtain best practices to identify how telehealth will exist in a post-covid era.

The initial transition to telehealth and COVID related reduction in in-person care resulted in overall fewer patient encounters. We explored evidence for increased provider workload through an analysis of EHR usage metadata comparing pre-covid (Mar, Apr, May 2019), mid-covid (Mar, Apr, May 2020), and late-covid (Mar, Apr, May 2021) time periods. EHR usage metrics do not comprehensively report on total provider work volume, they do accurately reflect the time and interaction providers have with the EHR. In a 2017 study published in the Annals of Family Medicine, the American Medical Association found that primary care physicians spend nearly 6 hours interacting with the EHR during and after clinic hours(3). Time in the EHR both total and afterhours can be significant and varied across provider sex (4). This study aims to evaluate our hypothesis that, in the wake of COVID-19 pandemic, clinician EHR time and clinical activities has increased.

Methods:

Cohort selection:

The selected cohort represented physician providers with ambulatory care time from all specialties and with EHR usage metrics in our Epic Signal report. There are a variety of clinician types in the Signal Report, but for this review we only focused on physician providers. Providers were excluded if they had no metric data for any of the selected time periods. A total of 1262 providers were clinically active with metric data for all three time frames.

Pre/Mid/Late COVID-19 Analysis Cohorts:

To evaluate changes in EHR usage we specifically split the selected cohort into pre-COVID, mid-COVID, and late-COVID groupings using mean values of EHR usage metrics across providers in general and by specialty. To account for seasonal variations of illness that can contribute to differences in clinical volume, the same three-month period of March/April/May for 2019, 2020 and 2021 were compared.

Signal Database:

The Epic Signal report summarized provider EHR usage metrics based upon user action log data. The Signal report divides provider activities in the EHR by seven categories: time in system, clinical review, note writing, ordering, in basket, in schedule and other. We identified a subset of metrics for our study (Table 1) derived in similar fashion as previously reported including calculated metrics of total messages, proportion of received messages completed, total
after hours time, and proportion of time after hours. We selected elements from each of these for inclusion in our study (Table 1).

<table>
<thead>
<tr>
<th>Table 1: Collected Attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Efficiency</strong></td>
</tr>
<tr>
<td>Avg Progress Note Length</td>
</tr>
<tr>
<td>Percent Appt Closed Same Day</td>
</tr>
<tr>
<td>Seconds per Completed Message</td>
</tr>
<tr>
<td>Message Turnaround Time</td>
</tr>
<tr>
<td><strong>Work Volume</strong></td>
</tr>
<tr>
<td>Total Appts</td>
</tr>
<tr>
<td>Total Received Messages</td>
</tr>
<tr>
<td><strong>Messages</strong></td>
</tr>
<tr>
<td>Total Completed Messages</td>
</tr>
<tr>
<td>Total In Basket Time</td>
</tr>
<tr>
<td>Proportion of Received</td>
</tr>
<tr>
<td>Time</td>
</tr>
<tr>
<td>Total Time in Clin Review</td>
</tr>
<tr>
<td>Total Time in Notes</td>
</tr>
<tr>
<td>Total Time in Orders</td>
</tr>
<tr>
<td>Total Time in Schedule</td>
</tr>
<tr>
<td>Total Time in System</td>
</tr>
<tr>
<td>Total Time After-hours</td>
</tr>
<tr>
<td>Proportion of Total Time</td>
</tr>
</tbody>
</table>

1\[
\text{These metrics are not directly reported in the signal database, but were calculated using signal metrics as defined above and previously described.}
\]

**Primary Outcome Assessment:** Our primary outcome measures included evaluation of provider work volume and time spent in the EHR. Work volume metrics were identified as to be the total number of appointments providers had per month and the number of messages received per month. Time metrics included total time spent in EHR clinical review, orders, in basket, and notes per month. Additionally, we looked at a total summative time spent in the EHR and time spent in the EHR after hours (outside of scheduled hours and on unscheduled days). To understand population level impact of COVID19 on EHR use, we averaged physician signal data over three month intervals (Table 1).

**Secondary Outcome Assessment:** We also looked at in-basket and provider efficiency metrics as secondary outcomes. In basket metrics included total number of messages completed by providers, total received messages, and calculated percentage of received messages that were completed. Efficiency metrics included average progress note length and percentage of appointments closed the same day. We assessed the cross specialty (specifically focusing on family medicine, endocrinology, pediatrics, psychiatry, and ophthalmology) absolute change in each of the aforementioned primary and secondary metrics (Table 1).

**Data Analysis:** Data analysis was conducted in R, an open-source data science software. Data was analyzed for basic descriptive statistics, basic counts, mean, and standard deviations for our data. Statistical significance was ascertained
utilizing unmatched two-way t-testing for differences in whole group pre, mid, and late data measures. Cross differences in terms of absolute in EHR usage metrics between the pre-mid and pre-late period were evaluated using one-way ANOVA.

_This study was approved by the Duke IRB Review Committee._

**Results:**

There were 1262 physician providers included in our analysis. These providers represented 66 specialties with the top 3 representative specialties being family medicine (n=149), internal medicine (n=94), and pediatrics (n=80). Additional specialty counts are included in Table 2.

<table>
<thead>
<tr>
<th>Table 2: Basic provider counts for primary and secondary analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Specialties (n)</strong></td>
</tr>
<tr>
<td><strong>Sub-analysis</strong></td>
</tr>
<tr>
<td>Family Medicine</td>
</tr>
<tr>
<td>Internal Medicine</td>
</tr>
<tr>
<td>Pediatrics</td>
</tr>
<tr>
<td>Cardiovascular Disease</td>
</tr>
<tr>
<td>Oncology</td>
</tr>
<tr>
<td>Ophthalmology</td>
</tr>
<tr>
<td>Orthopedic Surgery</td>
</tr>
</tbody>
</table>

_Full Analysis (Table 3):_

**Measures of work volume:** We evaluated two measures of work volume: total number of appointments and total number of messages received in the measurement period averaged over a month. The total average number of appointments per provider per month dropped significantly from Pre-COVID (mean(SD)=134(123)) to Mid-COVID (mean(SD)=93(86)), (Pre-mid p=<0.001) but volumes rebounded by comparison to Late-COVID (mean(SD)=130(117) values; (Pre-late p=0.404). The total number of messages received per provider per month initially dropped significantly (Pre-Covid(mean(SD)=459(380)) to Mid-Covid(mean(SD)=402(362)) (Pre-mid p-value=<0.001). Contrary to the relative return to Pre-COVID appointment volumes, the number of messages received significantly increased for the Late-COVID time period(mean(SD)=521(423); Pre-late p=<0.001).

_Time Spent in EHR:_ We collected data on provider time spent in the EHR for four activities (clinical review, in basket, orders, and in notes). There was a consistent trend demonstrating an initial decrease of provider total time spent in each EHR activity (refer to Table 3). This trend was followed by increases in time spent in EHR activities. These increases were significant in the case of time spent in clinical review, orders, and schedule (refer to Table 3). Total time spent in the system per provider per month, was significant for an initial drop in total time for the pre-mid reporting periods (Pre-COVID (mean(SD)=1958(1576)) Mid-COVID (1709(1473)) Pre-mid p-value=<0.001. With a return to baseline for total time spent in the EHR in the Late-COVID (2007(1563)) group. with no significant difference at the average provider level, Pre-late p-value=0.439).

_In-Basket Work:_ We evaluated several measures of in-basket specific provider work (completed message volume, average time spent on a message, average time between message receipt and message completion). Providers had an initial decrease, followed by subsequent increase in their total completed message volume (Pre-COVID (mean(SD)=416.69(360.96)) Mid-COVID (366.34(356.5)) Late-COVID (496.16(433.22)); Pre-mid p-value=<0.001, Pre-late p-value=<0.001), with an initial dip, but subsequent normalization of their efficiency in terms of seconds per completed message (Pre-COVID (mean(SD)=39.12(31.88) Mid-COVID mean(SD)=51.72(64.63) Post-COVID mean(SD)=40.56(34.92); Pre-mid p-value=<0.001, Pre-post p-value=0.283). Despite the increased volume, provider
response time had not significantly changed (Pre-COVID (mean(SD)=9.21(29.37) Mid-COVID mean(SD)=9.02(28.97) Post-COVID mean(SD)=9.46(36.78); Pre-mid p-value=0.868, Pre-post p-value=0.851).

**Efficiency and Complexity:** As a final measure of efficiency and complexity we evaluated provider documentation practices (for progress notes) and the percentage of provider encounter closures on the same day as the appointment. Providers had an initial significant decrease in progress note lengths, followed by subsequent significant increases in note lengths (Pre-COVID (mean(SD)=6912.48(3990.56)) Mid-COVID (6306.94(4127.15)) Late-COVID (7323.3(4262.38)); Pre-mid p-value=<0.001, Pre-late p-value=0.015) with no significant change in encounters closed the same day (Pre-COVID (mean(SD)=0.62(0.31)) Mid-COVID (0.63(0.32)) Late-COVID (0.62(0.33)); Pre-mid p-value=0.429, Pre-late p-value=0.967).

**Table 3:** Evaluated mean (standard deviation) values for measures pre COVID (pre), shortly after COVID (mid), and nearly 1 year after COVID onset (late)

<table>
<thead>
<tr>
<th></th>
<th>Pre (mean(sd))</th>
<th>Mid (mean(sd))</th>
<th>Late (mean(sd))</th>
<th>PreMid p-value</th>
<th>PreLate p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Efficiency</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Avg Progress Note Length</td>
<td>6913(3991)</td>
<td>6307(4127)</td>
<td>7323(4262)</td>
<td>&lt;0.001</td>
<td>0.015</td>
</tr>
<tr>
<td>Percent Appt Closed Same Day</td>
<td>0.62(0.31)</td>
<td>0.63(0.32)</td>
<td>0.62(0.33)</td>
<td>0.429</td>
<td>0.967</td>
</tr>
<tr>
<td>Seconds per Completed Message</td>
<td>39.1(31.9)</td>
<td>51.7(64.6)</td>
<td>40.6(34.9)</td>
<td>&lt;0.001</td>
<td>0.283</td>
</tr>
<tr>
<td>Message Turnaround Time</td>
<td>9.2(29.4)</td>
<td>9.0(29.0)</td>
<td>9.5(36.8)</td>
<td>0.868</td>
<td>0.851</td>
</tr>
<tr>
<td><strong>Work Volume</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total Appts</td>
<td>134.1(122.7)</td>
<td>92.5(86.5)</td>
<td>130.1(116.9)</td>
<td>&lt;0.001</td>
<td>0.404</td>
</tr>
<tr>
<td>Total Received Messages*</td>
<td>459.2(379.5)</td>
<td>400.5(361.6)</td>
<td>520.5(422.9)</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td><strong>Messages</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total Completed Messages</td>
<td>416.7(361.0)</td>
<td>366.3(356.5)</td>
<td>496.2(433.2)</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Total In Basket Time (min)</td>
<td>224.7(212.7)</td>
<td>237.9(238.3)</td>
<td>274.5(252.9)</td>
<td>0.144</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Proportion of Messages Completed</td>
<td>1.24(0.82)</td>
<td>1.31(0.81)</td>
<td>1.17(0.61)</td>
<td>0.021</td>
<td>0.031</td>
</tr>
<tr>
<td><strong>Time</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total Time in Clin Review (min)</td>
<td>311.4(272.3)</td>
<td>277.9(247.2)</td>
<td>350.0(293.7)</td>
<td>0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Total Time in Notes (min)</td>
<td>769.19(703.59)</td>
<td>624.55(619.17)</td>
<td>789.95(708.87)</td>
<td>&lt;0.001</td>
<td>0.464</td>
</tr>
<tr>
<td>Total Time in Orders (min)</td>
<td>198.08(241.91)</td>
<td>166.49(207.32)</td>
<td>240.81(279.32)</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Total Time in Schedule (min)</td>
<td>138.04(106.3)</td>
<td>136.84(114.92)</td>
<td>156.43(116.94)</td>
<td>0.785</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Total Time in System (min)</td>
<td>1958.2(1576.27)</td>
<td>1708.56(1472.69)</td>
<td>2006.55(1562.83)</td>
<td>&lt;0.001</td>
<td>0.439</td>
</tr>
<tr>
<td>Total Time Afterhours(min)*</td>
<td>811.75(717.45)</td>
<td>692.35(664.71)</td>
<td>847.48(706.53)</td>
<td>&lt;0.001</td>
<td>0.254</td>
</tr>
<tr>
<td>Proportion of Time Outside Scheduled Hours*</td>
<td>0.39(0.21)</td>
<td>0.36(0.19)</td>
<td>0.39(0.21)</td>
<td>&lt;0.001</td>
<td>0.566</td>
</tr>
</tbody>
</table>

**Specialty Analysis (Table 4):**

**Efficiency:** Our specialty sub-analysis revealed that many of the metrics differed by specialty. Average progress note length significantly differed across specialties both in the pre-mid and pre-late analysis (Pre-mid p-value=<0.001, Pr-
late p-value=<0.001), specialties also differed with respect to the total appointments closed the same day within our pre-mid analysis, but not for the pre-late analysis (Pre-mid p-value=0.008, Pre-late p-value=0.51).

**Work Volume:** Total appointments significantly differed by specialty in our pre-mid and pre-late analysis (Pre-mid p-value=<0.001, Pre-late p-value=<0.001). Total received messages, on the other hand differed for only the pre-mid analysis (Pre-mid p-value=<0.001, Pre-late p-value=0.087).

**Messages:** The proportion of completed messages did not differ by specialty in the pre-mid analysis, but did differ in the pre-late analysis (Pre-mid p-value=0.572, Pre-late p-value=0.043), total completed messages, on the other hand only differed across specialties for our pre-mid analysis (Pre-mid p-value=<0.001, Pre-late p-value=0.149), whereas total in-basket time differed only in our pre-late analysis (Pre-mid p-value=0.055, Pre-late p-value=<0.001).

**Time:** Specialties significantly differed by total time in orders (Pre-mid p-value=0.003, Pre-late p-value=0.05), total time in schedule (Pre-mid p-value=0.022, Pre-late p-value=0.002), and total time in notes (Pre-mid p-value=<0.001, Pre-late p-value=0.01), and proportion of time outside scheduled hours (Pre-mid p-value=0.002, Pre-late p-value=<0.001) for both the pre-mid and pre-late analysis. However, total time in system only significantly differed between specialties for our pre-mid analysis (Pre-mid p-value=<0.001, Pre-late p-value=0.068).

### Table 4: Cross Specialty Comparison of Changes in EHR Usage Metrics during the Pre-MidCOVID and Pre-LateCOVID Time Periods.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Comparison</th>
<th>Endocrinology</th>
<th>Family Medicine</th>
<th>Ophthalmology</th>
<th>Pediatrics</th>
<th>Psychiatry</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Appts</td>
<td>-9.39(20.18)</td>
<td>-57.12(62.19)</td>
<td>-136.11(108.22)</td>
<td>-73.55(79.05)</td>
<td>-8.57(19.04)</td>
<td>&lt;0.001</td>
<td></td>
</tr>
<tr>
<td>Percent Appt Closed</td>
<td>-0.02(0.24)</td>
<td>0.01(0.16)</td>
<td>0.06(0.23)</td>
<td>0.03(0.22)</td>
<td>-0.12(0.24)</td>
<td>0.008</td>
<td></td>
</tr>
<tr>
<td>Same Day</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Avg Progress Note Length</td>
<td>-4192.81(3284.79)-209.14(1349.89)</td>
<td>32.63(398.85)</td>
<td>-137.16(2004.31)-3365.25(4300.97)</td>
<td>&lt;0.001</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proportion of Time Outside Scheduled Hours*</td>
<td>-0.03(0.11)</td>
<td>-0.01(0.07)</td>
<td>0.01(0.1)</td>
<td>-0.02(0.11)</td>
<td>0.08(0.1)</td>
<td>0.002</td>
<td></td>
</tr>
<tr>
<td>Total Received Messages*</td>
<td>25.06(125.05)</td>
<td>-84.56(189.59)</td>
<td>-90.65(97.87)</td>
<td>-59.32(136.07)</td>
<td>10.31(107.38)</td>
<td>&lt;0.001</td>
<td></td>
</tr>
<tr>
<td>Seconds per completed message</td>
<td>(5.16(38.32))</td>
<td>(35.29(137.25))</td>
<td>(12.93(18.03))</td>
<td>(11.53(23.45))</td>
<td>(23.42(70.26))</td>
<td>0.242</td>
<td></td>
</tr>
<tr>
<td>Turnaround Time</td>
<td>(7.29(26.27))</td>
<td>(32.85(35.79))</td>
<td>(-1.55(48.45))</td>
<td>(-1.03(15.80))</td>
<td>(0.55(11.59))</td>
<td>0.683</td>
<td></td>
</tr>
<tr>
<td>Proportion of Messages Completed</td>
<td>(-0.03(0.37))</td>
<td>(0.01(0.43))</td>
<td>(0.07(0.32))</td>
<td>(0.01(0.40))</td>
<td>(-0.09(0.75))</td>
<td>0.572</td>
<td></td>
</tr>
<tr>
<td>Total Time in Notes (min)</td>
<td>-114.13(275.69)</td>
<td>-182.16(422.91)</td>
<td>-466.18(374.43)</td>
<td>-297.15(504.37)</td>
<td>-161.36(459.95)</td>
<td>&lt;0.001</td>
<td></td>
</tr>
<tr>
<td>Total Time Outside Sched Hours (min)*</td>
<td>-170.05(578.29)</td>
<td>-128.79(392.28)</td>
<td>-233.75(330.94)</td>
<td>-204.35(481.99)</td>
<td>188.95(344.93)</td>
<td>0.008</td>
<td></td>
</tr>
<tr>
<td>Total Time in Clin Review (min)</td>
<td>-35.65(152.18)</td>
<td>-66.06(141.6)</td>
<td>-22.98(87.2)</td>
<td>-62.6(117.94)</td>
<td>-15.93(123.41)</td>
<td>0.078</td>
<td></td>
</tr>
<tr>
<td>Total In Basket Time (min)</td>
<td>51.49(127.54)</td>
<td>22.33(128.68)</td>
<td>-8.33(55.21)</td>
<td>25.15(103.29)</td>
<td>55.62(118.25)</td>
<td>0.055</td>
<td></td>
</tr>
<tr>
<td>Total Time in Orders (min)</td>
<td>-37.18(61.24)</td>
<td>-88.42(152.92)</td>
<td>-64.37(55.59)</td>
<td>-80.44(92.55)</td>
<td>-1.95(65.32)</td>
<td>0.003</td>
<td></td>
</tr>
<tr>
<td>Total Time in Schedule (min)</td>
<td>0.87(37.1)</td>
<td>-5.05(105.73)</td>
<td>-45.31(80.75)</td>
<td>-8.36(103.92)</td>
<td>13.86(79.93)</td>
<td>0.022</td>
<td></td>
</tr>
<tr>
<td>Total Time in System (min)</td>
<td>-145.01(605.73)</td>
<td>-381.03(890.52)</td>
<td>-877.86(653.57)</td>
<td>-545.89(898.7)</td>
<td>-75.51(799.84)</td>
<td>&lt;0.001</td>
<td></td>
</tr>
<tr>
<td>Total Completed Messages</td>
<td>66.57(175.68)</td>
<td>-55.02(199.3)</td>
<td>-81.34(108.19)</td>
<td>-55.27(124.55)</td>
<td>3.12(104.68)</td>
<td>&lt;0.001</td>
<td></td>
</tr>
<tr>
<td>Table 4 Continued</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total Appts</td>
<td>5.83(26.6)</td>
<td>-31.68(68.25)</td>
<td>26.13(129.3)</td>
<td>-35.41(76.45)</td>
<td>10.02(30.89)</td>
<td>&lt;0.001</td>
<td></td>
</tr>
<tr>
<td>Percent Appt Closed Same Day</td>
<td>0.07(0.29)</td>
<td>0(0.2)</td>
<td>0.03(0.19)</td>
<td>0.02(0.23)</td>
<td>0(0.28)</td>
<td>0.51</td>
<td></td>
</tr>
<tr>
<td>Avg Progress Note Length</td>
<td>-561.86(1947.64)</td>
<td>276.44(1396.82)</td>
<td>187.57(623.28)</td>
<td>885.63(2298.57)</td>
<td>-1850.37(4088.26)</td>
<td>&lt;0.001</td>
<td></td>
</tr>
<tr>
<td>Proportion of Time Outside Scheduled Hours*</td>
<td>0.06(0.14)</td>
<td>0.01(0.08)</td>
<td>-0.01(0.08)</td>
<td>-0.02(0.12)</td>
<td>0.06(0.08)</td>
<td>&lt;0.001</td>
<td></td>
</tr>
<tr>
<td>Total Received Messages*</td>
<td>125(162.35)</td>
<td>17(283.39)</td>
<td>74.45(117.22)</td>
<td>33.6(190.31)</td>
<td>78.67(113.08)</td>
<td>0.087</td>
<td></td>
</tr>
<tr>
<td>Seconds per completed message</td>
<td>(-2.21(40.00))</td>
<td>(7.62(31.72))</td>
<td>(-1.78(14.21))</td>
<td>(8.05(32.17))</td>
<td>(-7.27(46.80))</td>
<td>0.043</td>
<td></td>
</tr>
<tr>
<td>Turnaround Time Proportion of completed messages</td>
<td>(-0.68(4.51))</td>
<td>(1.54(33.22))</td>
<td>(-3.78(48.02))</td>
<td>(-1.12(16.63))</td>
<td>(5.90(19.24))</td>
<td>0.673</td>
<td></td>
</tr>
<tr>
<td>Total Time in Notes Pre-Late Change (late-pre)</td>
<td>(-0.09(0.16))</td>
<td>(-0.05(0.23))</td>
<td>(0.04(0.31))</td>
<td>(-0.02(0.39))</td>
<td>(-0.23(0.98))</td>
<td>0.043</td>
<td></td>
</tr>
<tr>
<td>Total Time Outside Sched Hours(min)*</td>
<td>101.46(687.41)</td>
<td>14.44(409.29)</td>
<td>6.19(272.74)</td>
<td>-55.39(490.85)</td>
<td>218.49(306.57)</td>
<td>0.189</td>
<td></td>
</tr>
<tr>
<td>Total Time in Clin Review (min)</td>
<td>51.4(203.41)</td>
<td>37.41(168.98)</td>
<td>25.49(85.27)</td>
<td>27.81(153.55)</td>
<td>44.88(150.11)</td>
<td>0.926</td>
<td></td>
</tr>
<tr>
<td>Total In Basket Time (min)</td>
<td>151.22(204.63)</td>
<td>35.82(160.87)</td>
<td>14.25(54.35)</td>
<td>71.01(113.53)</td>
<td>89.96(148.21)</td>
<td>&lt;0.001</td>
<td></td>
</tr>
<tr>
<td>Total Time in Orders (min)</td>
<td>55.61(84.24)</td>
<td>69.82(200.43)</td>
<td>35.42(60.79)</td>
<td>8.98(109.38)</td>
<td>71.13(145.83)</td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td>Total Time in Schedule (min)</td>
<td>6.66(42.99)</td>
<td>3.45(71.99)</td>
<td>27.98(79.62)</td>
<td>3.49(91.23)</td>
<td>67.57(119.55)</td>
<td>0.002</td>
<td></td>
</tr>
<tr>
<td>Total Time in System (min)</td>
<td>138.56(863.5)</td>
<td>-149.73(988.09)</td>
<td>103.66(669.74)</td>
<td>-48.42(954.36)</td>
<td>317.39(931.99)</td>
<td>0.068</td>
<td></td>
</tr>
<tr>
<td>Total Completed Messages</td>
<td>147.45(171.7)</td>
<td>38.43(281.23)</td>
<td>67.61(125.89)</td>
<td>38.55(174.92)</td>
<td>75.4(126.59)</td>
<td>0.149</td>
<td></td>
</tr>
</tbody>
</table>
Figure 1. Total time in system, Messages Received, Total Afterhours Time for specialties pre, mid, and late COVID. (title to include total time in system per month)

Discussion:

We present one of the first assessments of the impact of the COVID-19 pandemic on changes in clinical activity as measured by provider EHR usage metrics. Our findings revealed an initial reduction in EHR related usage by time and volume metrics after COVID-19 lockdowns and overall decrease in clinical encounters. These reductions in EHR usage were not uniform across all specialties, with some specialties having significant increases in time and volume metrics. More than one year into the pandemic and with a loosening of COVID-19 restrictions, most EHR usage metrics have returned to baseline with the exception of increases for total messages received.

We found an initial decrease in provider volume during the months of March to May for total appointments and total time in the system. While these changes corresponded with reductions in EHR usage, the reductions were disproportionately modest. We hypothesize multiple factors accounted for these non-concordant changes. First, is the rise of telehealth. At our institution, telehealth encounters are initiated by the provider within the EHR. These encounters occur directly in the EHR with the patient as opposed to previous in-person formats where the provider usage of the EHR was utilized as needed. It may be that with providers initiating telehealth encounters within the EHR; they are more apt to chart during the clinical encounter. Telehealth encounters being in the EHR led to more overall time of use with the encounter. The specialty with the highest utilization of telehealth is our Department of Psychiatry. We observed our psychiatrists with an increase of 317 minutes of average total time spent in the system per month comparing pre to late covid. The implementation of telehealth and compensatory increase in EHR time may need further policy and oversight.

An additional instance where there is a potential for increasing time in the EHR is with our endocrinologists who manage diabetes and making changes remotely while reviewing labs in the EHR, especially the time in system per day (Pre(mean(SD)= 2435.28(1799.19), Mid(mean(SD)=2547.88(1927.34), Late(mean(SD)=3231.08(1960.24)). This may be reflective of the fact that many of the endocrinologists are managing chronic diseases, such as diabetes and thyroid disease that need persistent lab work. This poses a unique disease process that potentially contributed to the EHR usage of our endocrinologists. Surgical subspecialties such as ophthalmology were greatly impacted by the change in COVID-19 era practice changes. They initially saw a decrease in appointments per day PreMid(mean(SD)=2...
113.77(109.83) and total time in system (PreMid(mean(SD)=−766.23(629.15)). This is likely reflected in the nature of the subspecialty requiring in-person care.

Nevertheless, if these were the sole factors contributing to the reduction in time, then we would not also see an increase in the total time providers spent in the EHR outside scheduled hours. Rather, we believe that additional factors are at play. The change to predominantly asynchronous patient-provider interactions had significant changes in provider EHR usage. Shortly after the pandemic providers saw a 20.3% increase in the number of received patient messages, possibly partially accounting for those stable patients who opted to forego clinical visits. Interestingly, the short-term impact of COVID-19 on provider EHR usage was not uniform across specialties.

While most specialties saw a significant decrease in EHR total time usage shortly after initiating COVID-19 public health mandates, psychiatry saw a relatively small decrease in total time in the system. As the department rapidly transitioned to telehealth care they did not see a decline in total appointment volume partially due to a relative decrease in no show volumes. Emerging evidence demonstrating a significant impact of the COVID-19 pandemic on patient mental health(1), we hypothesize both the patients increased desire for mental health services as well as improved accessibility of services through telemedicine encounters. The psychiatry department has been able to uniquely transition the majority of their ambulatory care encounters to telemedicine formats that allow for providers to provide off site care. While this served to reduce contact risk, many have reported technology challenges for both the patients and the providers with the transition.

Consistent with reports in popular culture, providers working from home are finding themselves managing clinical work in what for some is possibly a more distracted environment(5). This too may have impacted provider EHR usage time.

Of notable concern were the increases in total messages received in our Late-COVID cohort nearly one year after the start of the pandemic along with the relative normalization back for EHR time metrics corresponding to resumption to more in person encounters. The reasons for many of these changes are likely explained by increased patient acceptance of asynchronous messaging in place of in person appointments along with mild compensatory increases in appointment volumes making up for previously deferred in person appointments.

Providers have not gone back to the pre-COVID status quo, in terms of clinical encounters or EHR usage. Our results suggest trends towards increased volume of work by evidence of increasing number of messages along with a similar volume of clinical encounters. Certainly, there has been increasing evidence demonstrating increased provider burnout during the pandemic(6). While provider EHR usage represents a sliver of the factors influencing that burden, it is one that requires addressing. If not addressed, implications of provider burnout could mean worsening patient care(7), decreased provider productivity(6), and increased organizational turnover(8).

While some may look at our data and take hope in our providers improved calculated efficiency metrics (with providers closing encounters in a timelier manner), we urge caution when evaluating these factors. If evaluated from the perspective of the Yerkes-Dodson law (which suggests bell-curve relationship between arousal-stress- and performance) it may be that our providers are at the cusp beyond which their performance would deteriorate. Furthermore, this heightened work-burden is likely to not be sustainable and we anticipate declining performance over time if such burden is to persist. With this ever-changing landscape there is an opportunity for policy to improve the implementation and reimbursement of these visits. There have been new evaluation and management (E/M) codes pertaining to time coverage, however clinicians are also using the EHR on unscheduled days. There is also a potential for improvement in automated transcription systems to have an impact on documentation.

**Limitations:** There are several limitations to our study. Firstly, our analysis is reflective of a single academic health system. Although we are unable to provide evidence for other health systems, this represents a large dataset of over 1200 providers from 66 specialties. Our state and institution did not experience an overwhelming surge in COVID19 patient volumes during the three-month time periods for 2020 or 2021. Consequently, our results are more reflective of our state and local pandemic public health constraints. Our institution constrained in person encounters and did not permit elective surgical procedures initially. Additionally, there may be variation in how healthcare systems responded to the arrival of COVID19 in regard to limited resources and COVID patient volumes.

Another limitation of this data is lack of granularity of visit type, specifically differentiation of telehealth encounters. Given the relative rapidity of significant use of these technology platforms, not having this in the Signal data is not
surprising. Future Signal data at the encounter type level would allow for greater exploration into the effect of telehealth on provider EHR usage data.

We must also consider the fact that the time spent in the EHR is not comprehensive of total clinical care. Provider time in front of their patients and interacting with them and other members of the clinical team is not captured in these metrics fully. The Signal EHR metric reports also only represents ambulatory EHR usage, inpatient encounters, where more COVID patients receive care is not represented fully.

This study did not provide a direct measurement of the impact of increased EHR metrics on our providers or our patients in terms of health-related outcomes. Further investigation is needed to see how increased EHR time may impact physicians or patient outcomes. We instead showed evidence from our usage data that there has been an increase in patient-initiated messages and time in the EHR.

Conclusion: We have demonstrated that there is an increasing trend in EHR usage. Further investigation is needed in how EHR usage has changed for other provider types and specialties. Both specialty and different provider characteristics may contribute uniquely in how the EHR is used and how they responded to the pandemic. Specialties that rely predominantly on telehealth, like the Psychiatry department at our institution, are more likely to have increased EHR usage. Further investigation is needed in how telehealth specifically impacts EHR usage, provider wellbeing, and quality of care is needed. This increase in EHR usage may also reflect a decrease in no-shows, increased provider efficiency, or related to patient convenience. The healthcare system will need to continue to adapt to new challenges to better serve its patients.

References


5. Roose K. Sorry, but working from home is overrated. NY Times . 2020;10.


Healthy Lifestyle and Mood: A Biomedical Informatics Citizen Science Project in a High School Classroom

Jennifer Ushe¹, Doug Redd, Ph.D.²,³, Scarlyn Gutierrez Nunez², Eduardo A. Trujillo-Rivera Ph.D.², Senait Tekle², Stuart J. Nelson, MD², Qing Zeng-Treitler, Ph.D.²,³

¹Alexandria City Public Schools, Alexandria, VA 22302 ²George Washington University, Washington, DC 20037 USA, ³Washington VA Medical Center, Washington, DC 20422 USA,

Abstract

Mental health is an increasing concern in adolescents. Mental health disorders can affect academic performance, affect the cultivation of healthy relationships, and even lead to suicide. Healthy lifestyle can improve mental health, though there are gaps in the research, partly resulted from the lack of detailed longitudinal datasets on lifestyle and mental health. To inform and engage students in the research on adolescent lifestyle and mood, the George Washington University and the T.C. Williams High School in Alexandria, Virginia teamed up in a citizen science project. Students generated questions, collected data on themselves, analyzed the data, and produced research reports relating to their mental health and lifestyle. Student feedbacks suggest that the students find the project to be generally interesting and some students (46%) reported that the participation in the project may influence their college and career plans. The anonymized dataset resulted from the project provides another contribution to science.

Introduction

There has been a rise in mental health issues in adolescents[1]. Adolescents are a unique population, going through changes in the way they think, feel, and interact with others. It is a normal part of development for teens to experience a wide range of emotions. Some teens, however, go on to have mental disorders. According to a US Department of Health and Human Services report, one in six adolescents (2.3 million/year) has had a severe mental health disorder at some point in their life[2]. Mental health disorders among adolescents can affect academic performance, affect the cultivation of healthy relationships, and even lead to suicide[3]. Particularly of concern is that rates of major depressive episodes increased by 52% between 2005 and 2017 among adolescents aged 12 to 17[1]. According to a 2019 publication, from the mid-2000s to 2017, there has been a sharp rise in the rates of suicidal thoughts, plans, and attempts.

At the same time, many adolescents do not adhere to a healthy lifestyle. Sedentary behavior continues to be a challenge, with less than 24% of children 6 to 17 years of age getting the recommended 60 minutes of physical activity daily[4]. In addition, most youth do not meet the recommended healthy eating habits[5]. Physical activity and nutrition deficits correlate with the increasing number of health issues in adolescents, including mood and weight problems. Between 1999-2000 and 2015-2016, United States obesity rates in youth increased from 13.9% to 18.5%[6]. There are many risks resulting from obesity, including high blood pressure, elevated cholesterol, breathing problems, and musculoskeletal pain. Obesity is also linked to lower self-reported quality of life, social problems, and even anxiety and depression[7].

However, gaps remain in the research on the interaction between adolescents’ lifestyle and mental health. One big challenge is the availability of a detailed, longitudinal dataset on lifestyle and mental health. For those datasets that do exist, the populations are often based in other countries, including England, Canada, and France, and the data tended to be collected from cross-sectional studies[8-12]. In addition, existing datasets still lack nuanced data to study the complex relationship between lifestyle and mood. One example is Add Health, the largest longitudinal study on adolescent health that began in 1995 and has undergone five waves of data collection, the most recent in 2016[13].

1169
While this comprehensive data set is impressive, there were 2-8-year intervals between waves of data collection, and adolescence is a period of rapid changes.

Another challenge, perhaps more social or cultural, is to engage adolescents in the enhancement and management of their health through research. Recruitment of subjects for clinical research is often not easy due to several factors: There may not be any immediate benefit from participation. Today's adolescents often have a packed schedule. Asking adolescents to track their lifestyle and mood closely over a period of time confers some burden and inconvenience.

Citizen science in the classroom is a new approach that has the potential to benefit students and to further understanding. The rise of “citizen science” offers an alternative to the traditional paradigm that separates researchers from the subjects they study. In citizen science studies, scientific work is undertaken by members of the general public, often in collaboration with or under the direction of professional scientists and scientific institutions[14-16]. In our citizen science project, students serve as both the study subjects and researchers, both as data source and as data analyst. They formulate research questions of interest to themselves and observe the impact in their own life. Collecting this data and building a resultant, anonymized dataset provides another contribution to science.

Specifically, to inform and engage students in the research on adolescent lifestyle and mood, the George Washington University (GW) School of Medicine and Health Sciences Biomedical Informatics Center and T.C. Williams High School in Alexandria City Public Schools (ACPS) Virginia teamed up in a citizen science project. Since the fall of 2016, GW and ACPS have collaborated on various projects. This collaboration has evolved into a Governor’s Health Sciences Academy and sophisticated in-class research projects. During the 2018-2019 school year, the students' research project took on a citizen science approach, developing the research questions to ask, and serving as subjects of the study. Based on input from the teacher and students, healthy lifestyle and mood were selected as the focus.

Background

The T.C. Williams High School is located in Alexandria City, where the median household income is $93,370, and 43% of households contain a married couple[17]. T.C. Williams High School is the largest high school in Northern Virginia, with 3,959 students between two campuses. The demographics of T.C. Williams HS is very diverse: Hispanic (40.8%), African American (28.2%), and White (24.1%)[18].

Mental health and obesity are prevalent health issues in Alexandria City. Between Spring 2018 and Summer 2019, the Alexandria Health Department conducted a Community Health Assessment, gaining insight from local data and community input regarding various health issues. One in five adults (21%) reported poor mental health lasting more than five days, and 22% of adults reported being obese, with community members reporting mental health and obesity as two of the top five health-related issues in Alexandria City[19]. These assessment results are consistent with national findings and support the need for further research on adolescent lifestyle choices and their relationship to mood. Moreover, compared to the national average of 49%, only 31% of Alexandria High School students reported regular exercise.

Methods

Students in the Biotechnology course at T.C. Williams High School, assisted by T.C. Williams and GW's faculty, identified pressing health issues, conducted literature research, created research questions and hypotheses, and performed data analysis submitted a poster presentation for the course.

IRB approval was requested and approved by the GW. Consent forms were sent to all parents, requesting permission for their son or daughter to participate. Additionally, students were given the option to opt-out of the data collection part of the study. A total of 3 students did not participate in the data collection but participated in other parts of the study. The remaining 18 female (64.3%) and 10 male (35.7%) students were provided an external study ID for data collection. All students were between the age of 16 and 17.

Students organized themselves into groups of two to five students in which they were to complete their research. At the beginning of the project, students were provided information on what data would be collected, allowing each group to determine a research focus. Once a group had chosen a focus, they worked closely with GW staff and ACPS library staff to carry out the research.

Students conducted literature reviews using T.C. Williams High School’s library resources with school librarians' guidance. Library resources utilized included databases, such as Science in Context and PubMed, Google Scholar, and books. After completing a search for relevant journal articles, students were instructed on how to complete a journal article review and create an American Psychological Association (APA) citation.
In addition to journal reviews, Children’s National Medical Center experts were invited to meet with the students. Two public health and policy professionals, as well as a medical doctor, spoke to the students about brain development and the current understanding of the effects of nutrition, sleep, and activity level on adolescents’ mood.

The GW staff taught a short lesson on how to develop a hypothesis and research questions. This lesson included examples of research questions, the importance of conducting a literature review before creating a research question, and the difference between null and alternative hypotheses.

To introduce the concept of research questions, students were shown a sample case study and asked to formulate two related questions. Students then categorized these questions into background and foreground questions. Once they had an understanding of the types of research questions, student groups were tasked with brainstorming potential research questions related to the study. Students were asked to follow a 5-step process to create their own research question for the project, which required them to think about the topic of interest, ideas and issues they would want to explore, problems within that topic of choice, questions they have based on their literature reviews and the context in which the study would be conducted.

Data collection was conducted between December 2018 - January 2019. Study participants completed a one-time questionnaire, daily mood tracker for thirty days, and fitness trackers to track their daily steps, physical activity, and sleep.

They were first instructed to complete, on paper, the background questionnaire. The background questionnaire was designed using previously tested and validated survey questions. It contained a series of questions regarding how students spend their time outside of school, including what activities they engaged in and how much time per week was spent on each activity. Additionally, at the students' request, the questionnaire included time spent on electronics (for personal and educational use) per day. The questionnaire was administered during the biotechnology class via paper format.

Consented students were then provided with Garmin Vivofit 3[20] fitness trackers and instructed on proper usage, including how and when to wear them. Students were asked to wear the tracker beginning December 14, 2018, through January 14, 2019, and only remove it when they were in the water, competing in activities that restricted its use, and when it was charging. It was recommended to students that they charge their device on Sunday and Wednesday evenings. At the end of the data collection period, students returned their trackers so that data (both sleep and number of steps) could be compiled.

In addition to collecting activity level and sleep data, students were given access to the Daily Nutrition and Mood Tracker, a questionnaire administered using SurveyLegend, a web-based service. Students were instructed to set an alarm to serve as a reminder to take the survey each evening before they went to bed. The daily nutrition and mood survey was completed each evening beginning Saturday, December 15, 2018, and ending Monday, January 14, 2019 (inclusive), for a total of 31 days.

The daily nutrition and mood tracker, created specifically for this study, served as a tool for students to log the food consumed at each of their meals (breakfast, lunch, dinner, and snacks) and indicate when each meal was consumed, exclusive of snacks. Additionally, students were asked to track their mood for three separate times of the day: morning (before 12 pm), afternoon (12-4 pm), and evening (after 4 pm). A numerical rating scale was used where one star indicates a poor mood and 5 stars, an excellent mood. Students were also asked to answer questions regarding the previous nights’ sleep, as well as indicate how many glasses of water they consumed each day.

Students worked closely with GW staff to organize the data in a format that would be useful to their research questions. The first step in data analysis was to have students describe the data through visuals. Next, students conducted a univariate analysis to determine any patterns in the data. This was done using Google Sheets on student Chromebooks. Students learned how to sort and filter data, count rows, and create simple graphs and charts.

Following this study, students were asked to complete an evaluation of the impact of the citizen science project and the collaboration with George Washington University. Students were asked whether they found the following topics interesting or boring: data, biomedical informatics, research, collaborating with GW and creating graphs. Information was submitted on what the students enjoyed most and least and whether this project played a role in their college and/or career plans.

Results

Student Characteristics
Besides the basic demographic information, students reported how many Advanced Placement (AP) and/or Dual Enrollment (DE) classes they were taking during the 2018-2019 school year, as well as indicate their involvement in various activities. Additionally, students reported how much time they spent per day on their phones, watching TV, and on their computers. A summary of these results can be found in Table 1 below.

**Table 1. Characteristics of Study Participants.**

<table>
<thead>
<tr>
<th>Characteristics of Study Participants</th>
<th>(N=28)</th>
<th>N (Total %)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>Male</td>
<td>10 (35.7%)</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>18 (64.3%)</td>
</tr>
<tr>
<td>Race/Ethnicity</td>
<td>Multi-Racial</td>
<td>1 (3.6%)</td>
</tr>
<tr>
<td></td>
<td>Asian</td>
<td>3 (10.7%)</td>
</tr>
<tr>
<td></td>
<td>Black</td>
<td>3 (10.7%)</td>
</tr>
<tr>
<td></td>
<td>White</td>
<td>21 (75%)</td>
</tr>
<tr>
<td>Number of Advanced Placement</td>
<td>1-2</td>
<td>7 (26.9%)</td>
</tr>
<tr>
<td>and/or Dual Enrollment Classes*</td>
<td>3-4</td>
<td>14 (53.8%)</td>
</tr>
<tr>
<td></td>
<td>5 or more</td>
<td>5 (19.2%)</td>
</tr>
<tr>
<td>Time Spent on Phone (per day)</td>
<td>&lt; 1 hour</td>
<td>0 (0%)</td>
</tr>
<tr>
<td></td>
<td>1-2.5 hours</td>
<td>9 (32.1%)</td>
</tr>
<tr>
<td></td>
<td>3-4.5 hours</td>
<td>12 (42.9%)</td>
</tr>
<tr>
<td></td>
<td>&gt; 5 hours</td>
<td>7 (25%)</td>
</tr>
<tr>
<td>Time Watching TV (per day)**</td>
<td>&lt;1 hour</td>
<td>15 (55.6%)</td>
</tr>
<tr>
<td></td>
<td>1-2.5 hours</td>
<td>11 (40.7%)</td>
</tr>
<tr>
<td></td>
<td>3-4.5 hours</td>
<td>1 (3.7%)</td>
</tr>
<tr>
<td></td>
<td>&gt;5 hours</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Time Spent on Computer for School Work (per day)</td>
<td>&lt;1 hour</td>
<td>0 (0%)</td>
</tr>
<tr>
<td></td>
<td>1-2.5 hours</td>
<td>10 (35.7%)</td>
</tr>
<tr>
<td></td>
<td>3 - 4.5 hours</td>
<td>10 (35.7%)</td>
</tr>
<tr>
<td></td>
<td>&gt;5 hours</td>
<td>8 (28.6%)</td>
</tr>
<tr>
<td>Time Spent on Computer for Personal Use (per day)</td>
<td>&lt;1 hour</td>
<td>15 (53.6%)</td>
</tr>
<tr>
<td></td>
<td>1-2.5 hours</td>
<td>8 (28.6%)</td>
</tr>
<tr>
<td></td>
<td>3-4.5 hours</td>
<td>1 (3.6 %)</td>
</tr>
<tr>
<td></td>
<td>&gt;5 hours</td>
<td>4 (14.3%)</td>
</tr>
<tr>
<td>Extracurricular Activities^</td>
<td>School Clubs</td>
<td>18 (64.3%)</td>
</tr>
<tr>
<td></td>
<td>Academic Organizations</td>
<td>15 (53.6%)</td>
</tr>
<tr>
<td></td>
<td>Recreational Sports</td>
<td>9 (32.1%)</td>
</tr>
<tr>
<td></td>
<td>Church Youth Groups</td>
<td>6 (21.4%)</td>
</tr>
<tr>
<td></td>
<td>After-School Work/Internship</td>
<td>12 (42.9%)</td>
</tr>
<tr>
<td></td>
<td>School Sponsored Sports</td>
<td>16 (57.1%)</td>
</tr>
</tbody>
</table>

* Number of AP/DE Classes - Missing 2  
** Time Watching TV - Missing 1  
^ Percentages will not add to 100 because students could select more than one activity

**Question Formulation**

The student groups generated a total of 7 research questions (Table 2). Due to an unexpected technical difficulty in downloading the fitness data in time for student analysis, one group revised their fitness question to focus on nutrition.
Table 2. Research questions.

<table>
<thead>
<tr>
<th>Mealtimes</th>
<th>Nutrition</th>
<th>Sleep and Mood</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Dinner time</td>
<td>4. Will a high variety of breakfast foods increase adolescents’ morning mood?</td>
<td>6. What is the impact of winter break on adolescents’ sleep and mood?</td>
</tr>
<tr>
<td>2. Breakfast time</td>
<td>5. Is there a relationship between a high protein diet and adolescents’ sleep quality?</td>
<td>7. How do advance level classes impact sleep and mood in high school juniors and seniors?</td>
</tr>
<tr>
<td>3. Lunch time</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Although the sample size for this study was small, students still noted some interesting observations. These findings are noted below in Table 3.

Table 3. Summary of findings by research question.

<table>
<thead>
<tr>
<th>Research Question</th>
<th>Findings</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>There was a slight increase in the mood when meals were consumed between 8-10 am (breakfast), 12 - 2 pm (lunch), and 4-6 pm (dinner) compared to breakfast before 8 am or after 10 am, lunch before 12 pm, or after 2 pm, and dinner before 4 pm, or after 6 pm.</td>
</tr>
<tr>
<td>2</td>
<td>Students who reported skipping breakfast had the worst morning mood compared to students who ate any breakfast.</td>
</tr>
<tr>
<td>3</td>
<td>Eating dinner after 8 pm resulted in a lower sleep score (measured by difficulty of falling and staying asleep and overall amount of sleep) as compared to those who ate dinner between 4-6 pm.</td>
</tr>
<tr>
<td>4</td>
<td>A majority of students reported only eating grains for breakfast. These students reported a better mood compared to students who ate grains &amp; dairy, only vegetables, dairy, and protein, and those that skipped breakfast. Therefore, no evidence was found that suggest that a high variety of breakfast food increase adolescent’s morning mood in this cohort.</td>
</tr>
<tr>
<td>5</td>
<td>The students did not find clear evidence for the existence of relation between a high protein diet and sleep quality.</td>
</tr>
<tr>
<td>6</td>
<td>Student mood was found to be better during winter break as compared to during the school year. Students also reported slightly more sleep during winter break than during the school year.</td>
</tr>
<tr>
<td>7</td>
<td>All the students included advance level classes in their curricula. It was not possible to evaluate this hypothesis.</td>
</tr>
</tbody>
</table>
Outcomes of Citizen Science Project

As a culminating task, students created posters to present at the annual Leveraging Big Data Student Showcase at T.C. Williams High School. Students were required to include the following elements on their posters: Background, Purpose, Methods, Results, Discussion, and Future Work. Attendees at the showcase included ACPS and GW Staff, parents, community members, and T.C. Williams HS teachers and students.

One of the greatest benefits of the collaboration was the opportunity for students to extend their learning outside the classroom. During the summer of 2019, five current and former T.C. Williams High School students completed an internship at the GW Biomedical Informatics Center. During this internship, students organized the current study data and collected and cleaned additional data from Health Sciences Academy students during a summer bridge program. In a final step, one of the students completed a data analysis, wrote an extended abstract, and submitted the abstract to the American Medical Informatics Association (AMIA) High School Scholars Symposium.

Evaluation

Twenty-eight students completed the evaluation survey and, of these, 13 students (46.4%) stated that this project may have played a role in their future college and career plans. Additionally, students were asked to report their interest level on various topics on a scale of 1 (boring) to 5 (interesting). All topics reached an average rating of greater than 3.5, suggesting that the students found the topics to be more interesting rather than boring. Table 4 includes the descriptive statistics of these interest levels.

Table 4. Student interest level in various topics.

<table>
<thead>
<tr>
<th>Topics</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data collection</td>
<td>3.85</td>
<td>0.76</td>
</tr>
<tr>
<td>Biomedical Informatics</td>
<td>3.89</td>
<td>0.99</td>
</tr>
<tr>
<td>Research in general</td>
<td>3.82</td>
<td>0.95</td>
</tr>
<tr>
<td>Collaborating with GW</td>
<td>4.11</td>
<td>1.03</td>
</tr>
<tr>
<td>Statistical analysis</td>
<td>3.68</td>
<td>1.16</td>
</tr>
<tr>
<td>Creating Graphs</td>
<td>3.61</td>
<td>1.03</td>
</tr>
</tbody>
</table>

It was also important to learn what students liked most and least about this project. Students were able to choose all tasks that applied and could also input any additional items about the project that they liked or disliked. The top four “likes” and “dislikes” are shown in Table 5 below.

Table 5. A summary of student likes and dislikes in regards to the project. (N - total %)

<table>
<thead>
<tr>
<th>What Students Enjoyed Most</th>
<th>What Students Enjoyed Least</th>
</tr>
</thead>
<tbody>
<tr>
<td>Working with GW Staff – 19 (67.9%)</td>
<td>Analyzing data – 10 (35.7%)</td>
</tr>
<tr>
<td>Working in a group – 16 (57.1%)</td>
<td>Creating a poster – 8 (28.6%)</td>
</tr>
<tr>
<td>Analyzing data – 16 (57.1%)</td>
<td>Conducting background research – 7 (25%)</td>
</tr>
<tr>
<td>Collecting data – 12 (42.9%)</td>
<td>Collecting data – 7 (25%)</td>
</tr>
</tbody>
</table>

It is clear from this feedback, that students enjoyed working on various aspects of this project. While students enjoyed the experience, they did provide meaningful and constructive feedback. A few statements from the high school students, taken from the evaluation, are shown below:

“I do not have any additional feedback on the project. I think that it was a useful learning experience. However, due to the limited time and sample size of the project, I feel that it was unable to really be used as an academic study; due to many outliers that were caused by that limited amount of data.”

“I loved this project. I think we should have collected more data because analyzing the data was hard.”

“I think that before sending the survey for approval, the staff should ask students to submit extra survey questions.”
“This project would have been more interesting if we had collected data over a longer period of time.”

Discussion

The Greek ideal of a sound mind in a sound body is supported by a good amount of modern evidence. It is to be noted that there is increasing evidence that physical activity can help improve academic performance, including cognitive skills and mood[21]. Past studies have also concluded that risky lifestyle activities, including inadequate exercise and poor diet, may lead to adverse health outcomes and chronic diseases in adulthood[22-24]. We recognize that bad mood is not the same as mental health issues. At the same time, mental health disorder requires a clinical diagnosis while mood can be tracked by individual who are diagnosed with a mental health disorder as well as those who are not. Individuals with mental health problems also often have mood issues.

In this citizen science project, a group of high school students collected lifestyle and mood data on themselves. They generated research questions, analyzed data, and produced research reports on various topics, including the effects of skipped meals, winter break on mood, the relationship between protein consumption and quality of sleep, and Advanced Placement (AP) courses’ impact on sleep and mood. Five students continued their work with the GW Biomedical Informatics Center as interns collecting additional data from a cohort of 100 students enrolled in the Governor’s Health Sciences Academy during the Academy’s 2019 summer bridge program. One of these interns, a rising senior at T.C. Williams High School, used this data to prepare a manuscript for submission to the American Medical Informatics Association (AMIA) High School Scholars Symposium.

As public school budgets continue to decrease and more emphasis is placed on meeting standards, it is important for teachers and school districts to be creative about how they engage students in scientific research. For students to understand the importance of the research, they must first understand how it is relevant to their lives and communities. Because of their focus on community-driven research, citizen science projects can be an excellent tool to engage students in these efforts. However, according to Mueller, Tippins, and Bryan, many citizen-science projects do not provide citizens the opportunity to see scientists in action, nor do they involve collaborating with these scientists to create larger-scale scientific studies. Additionally, citizens do not typically ask their own questions or use the data to construct analyses[25].

This project engaged high school students and teachers with informaticians, public health students, medical professionals, and government affairs personnel. Unlike most citizen science projects, this project allowed students to work directly with professionals in the field to ask their own questions, develop project protocol, and analyze their own data.

While there may be concerns regarding student-collected data, studies have shown that youth, when advised by adults, are capable of collecting reliable data[25]. In his book, The Shame of a Nation, Jonathan Kozol claims that young people can be trusted for the quality of their data because they approach issues without previous experiences, which might give them a bias[26]. Therefore, we cannot discredit the food, sleep, and mood logs submitted by these students. We argue this data is more accurate because students have a vested interest in learning about how their choices impact their well-being.

Student responses were generally positive toward this project. However, this research was not without its limitations. First, the sample size was small; 28 students began the project, and only 25 submitted any data. In addition, only 15 students (60%) tracked data for more than 20 (out of 31) days. Second, because of technical issues (the Garmin activity trackers only keeps two weeks locally while data was not automatically uploaded to a server), Garmin activity and sleep data were excluded from this study. At the end of year evaluation, 11 (39.3%) students reported they were initially interested in working with the Garmin sleep data and 17 (60.7%) students with the Garmin activity data. Unfortunately, because of the lack of Garmin data, students were unable to research whether activity level and/or sleep play any role in an adolescent’s well-being. Third, nutrition data gathering was challenging for many reasons. Although students were asked to be detailed in their responses, not all students followed these instructions, which made it difficult to quantify the amount of food consumed. Additionally, data had to be cleaned in such a way that prevented the identification of individual students, so foods were categorized based on traditional food groups. This task was not always as straightforward as it might appear. Lastly, and most importantly, was the lack of student knowledge in techniques of data analysis. Because this project was embedded into an existing course, there was limited time to instruct students on how to properly analyze data.

Conclusions and future work

Regardless of the limitations, this project proved to be extremely successful and worthy of future work. More work will need to be done to improve data collection to ensure a more robust and reliable dataset. In order to fully investigate the factors that impact adolescent well-being, it will be important to fix technical issues with the chosen fitness trackers, improve the process for nutrition tracking, and spend more time with students on data analysis. Additionally, after reviewing existing data sets, like Add Health, it may be worthwhile to expand the baseline survey to include
questions about family, friendships, and community. These improvements and additions will allow for a better understanding of the impact of lifestyle choices and factors on adolescent mood.

In the future, we plan to expand this data set locally, within T.C. Williams High School, and then with another high school in neighboring Prince George’s County, Maryland. (Current was underway in the 2019-2020 school year but cut short by the COVID pandemic.) Beyond that, developing a more robust data collection method that will support high school student citizen science and the collection of a much larger dataset is envisioned.

References

18. T.C. Williams demographics: T.C. Williams High School; [19
20. Garmin vivofit 3, activity tracker with 1+ year battery life, sleep monitoring and auto activity detection: Amazon; [20


A Practical Approach for Monitoring the Use of Copy-Paste in Clinical Notes

David K. Vawdrey PhD1,2, Casey Cauthorn MIE1, Diane Francis1, Kathy Hackenberg RN1, Gerald Maloney DO3, Benjamin A. Hohmuth MD, MPH1,3

1Geisinger Steele Institute for Health Innovation, Danville, PA; 2Columbia University Department of Biomedical Informatics, New York, NY; 3Geisinger Department of Medicine, Danville, PA

Abstract

The use of copy-paste in authoring clinical notes has been widely embraced by busy providers, but inappropriate copy-paste has been lambasted by critics for introducing risks related to patient safety and regulatory compliance. At an integrated academic health system with over 4,100 providers writing notes, we developed a pragmatic approach to assess the use of copy-paste. From January 1–December 31, 2020, approximately 2.3M inpatient notes and 6.6M ambulatory clinic notes were authored in our electronic health record. Of the inpatient notes, 42% used copy-paste, and 19% of overall note content was copied; in ambulatory notes, 18% used copy-paste and 12% of note content was copied. We describe an approach for including providers’ copy-paste usage statistics into the ongoing professional practice evaluation process required for hospital accreditation, thereby offering individual training opportunities related to the lack of use of copy-paste or its potential overuse.

Introduction

Given the challenges that contribute to provider burnout, including increasing documentation requirements and poor electronic health record (EHR) usability,1,2 copy-paste is a welcome function that allows clinicians to author more comprehensive notes more quickly. Recognizing the widespread use of copy-paste in EHRs and the associated patient safety and compliance risks, the Partnership for Health IT Patient Safety provided four recommendations aimed at health information technology vendors and healthcare delivery organizations: “1) Provide a mechanism to make copy and paste material easily identifiable; 2) Ensure the provenance of copy and paste material is readily available; 3) Ensure adequate staff training and education; 4) Ensure copy and paste practices are regularly monitored, measured, and assessed.” While the first two recommendations have become standard functionality in EHR systems such as Epic Hyperspace (Epic Corp., Verona WI), the others—staff training/education, and regular monitoring of copy-paste practices—are more difficult to adopt.

At Geisinger, an integrated academic health system with over 4,100 providers on the medical staff, we developed a pragmatic approach to assess the use of copy-paste in aggregate, and also provide individual training opportunities by including providers’ copy-paste usage statistics into the ongoing professional practice evaluation (OPPE) process that is required for hospital accreditation.

Background

Copy-Paste Concerns

The “note bloat” phenomenon, where clinical notes balloon in length as large amounts of nonessential information are inserted, has been decried by experts as both annoying and dangerous.3 In addition to the potential for harm that arises when notes contain inconsistent and outdated information, note bloat may result in salient information becoming buried or lost. Wrenn and colleagues found that more than half of the text in EHR progress notes originated from prior notes accounts,4 and a 2017 study by Wang and colleagues found that 82% of text in inpatient progress notes was copied or generated by a template.5

The complex payment and medicolegal environment in the U.S. is likely a primary driver of clinical note length. In a study of U.S. and non-U.S. hospitals using the same EHR, Dowling and colleagues found that ambulatory notes in the United States are 4x longer than notes written in other countries.6 Clinicians in the U.S. may use tools like copy-paste to more efficiently author notes that meet the real or perceived requirements of stakeholders from billing, compliance, and legal departments.
In 2014, the Office of the Inspector General for CMS released a report titled “CMS and Its Contractors Have Adopted Few Program Integrity Practices to Address Vulnerabilities in EHRs.” The report warned that “when doctors, nurses, or other clinicians copy-paste information but fail to update it or ensure accuracy, inaccurate information may enter the patient’s medical record and inappropriate charges may be billed to patients and third-party health care payers. Furthermore, inappropriate copy-pasting could facilitate attempts to inflate claims and duplicate or create fraudulent claims.”

The OIG report expressing concern about copy-paste did not lead to CMS standards about the practice; indeed, compliance risks related to copy-paste seem to be significantly diminished as a result of the changes CMS included in the Physician Fee Schedule Final Rule in 2020. In this Rule, CMS explains that “the physician, the PA [physician assistant], or the APRN [advanced practice registered nurse] who furnishes and bills for their professional services” is only required “to review and verify, rather than re-document, information included in the medical record” by physicians, residents, nurses, students or other members of the medical team” (emphasis added).

Notwithstanding the 2020 CMS changes related to documentation practices, liability risks exist with improper use of copy-paste. Copy-paste is an easy target for critics because of egregious “errors of commission,” where a copied block of text is repeated in a note indicating, for example, that a patient is ‘post-op day 2’ when it has been a week since the operation. Similar examples have been identified when a note describes a patient as having a different sex or age than what is recorded elsewhere in the medical record. No matter how infrequent these cases, they are memorable examples of the risk of improper use of copy-paste.

**Copy-Paste Benefits**

While concerns related to copy-paste have been widely reported for the past 15 years, the benefits of copy-paste have been less publicized. With numerous studies documenting that clinicians spend more time in the EHR than face-to-face time with patients and growing concerns about provider burnout, any tool to improve efficiency of note-writing or any policy change that reduces documentation burden is welcome. Hilliard and colleagues recently reported on EHR factors that were associated with clinician burnout. Among “efficiency” variables—such as use of a “Chart Search” function, the number of “SmartPhrases” (personalized text templates), and the use of ordering shortcuts—the only variable independently associated with decreased reporting of burnout was a clinician’s use of copy-paste.

While unchecked use of copy-paste can result in problematic note bloat, appropriate copy-paste can improve the comprehensiveness of notes and help providers avoid the invisible error of omission. For example, a conscientious provider might bring forward important clinical information using copy-paste such as “8mm adrenal mass requires errors of commission (emphasis added)." While these links could either pull information from other places in the EHR (e.g., past medical history, lab test results) or they could prepopulate blocks of text that were created by individual providers or at the system level. When reviewing notes in the EHR, the mechanism by which text was entered into the note (manually, via template, or via copy-paste) could be visualized as shown in Figure 1.

**Methods**

Geisinger is an integrated health system with 9 hospitals and over 200 ambulatory care sites located primarily in central and northeast Pennsylvania. Geisinger implemented the Epic EHR beginning in 1996. Copying of text in the EHR could be done in several ways. First, text could be selected, copied, and pasted between notes using the same commands one would use in any text editor or word processing application. The Epic EHR also provided a copy-forward function that allowed copying a prior note (e.g., from the previous day of an inpatient encounter) and modifying it as needed for the current day. For the purposes of this study, copy-paste refers to these two methods.

There are other EHR capabilities for autogenerating text that we did not evaluate in this study. For example, text could be inserted into notes automatically using links that were either built into a note template or invoked by a text command known as a “dot phrase.” These links could either pull information from other places in the EHR (e.g., past medical history, lab test results) or they could prepopulate blocks of text that were created by individual providers or at the system level. When reviewing notes in the EHR, the mechanism by which text was entered into the note (manually, via template, or via copy-paste) could be visualized as shown in Figure 1.
Figure 1. Example of how the EHR allowed the source of text in a clinical note to be identified. Based on the user’s selection, only text that has been copied into the note is highlighted; text entered manually or via template is shown in light gray.

We assessed copy-paste practices among providers authoring inpatient and ambulatory clinic notes between January 1, 2020 and December 31, 2020. Providers were grouped into departments and departments at our institution were grouped into the following Institutes: Anesthesia/Surgical Services, Cancer, Diagnostic Medicine, Heart, Medicine, Neuroscience, Orthopedics, Surgery, Women & Children’s. We measured providers’ use of copy-paste by quantifying the percentage of notes that included text content added via copy-paste. Content was considered copied if it was copy-pasted from another note, not entered using templates that populated boilerplate text or patient-specific text (e.g., medications, laboratory test results, vital signs, etc.)

We measured the average length of notes (in characters) written by providers in various clinical institutes, and we calculated “Average % Copied Content” by dividing the number of copied characters by the total characters in each note.
Integrating Copy-Paste Statistics into the Ongoing Professional Practice Evaluation Process

The Joint Commission requires hospitals to conduct Ongoing Professional Practice Evaluation (OPPE) for all practitioners who are granted privileges as part of their Medical Staff. The goal of OPPE and its companion process—Focused Professional Practice Evaluation (FPPE)—is to support early detection and response to performance issues that could negatively impact patient outcomes. To conduct effective OPPE, an objective, data-driven foundation for making re-privileging decisions is needed. At our institution, department chairs or their designees (such as vice chairs or division chiefs) had responsibility for reviewing data twice per year with each provider in their department and recommending continued/limited/denied privileges for each.

We added the following copy-paste metrics to our OPPE:

1) Total Ambulatory Notes Authored
2) % of Ambulatory Notes Containing Any Copied Content
3) Average % of Text in Ambulatory Notes That Was Copied
4) Total Inpatient Notes Authored
5) % of Inpatient Notes Containing Any Copied Content
6) Average % of Text in Inpatient Notes That Was Copied

At our institution, OPPE reports were created using the Midas Statit Physician Profile & Review tool (Conduent Care Management Inc., Tucson, AZ). Profiles (reports) were customized based on the specialty of each physician/advanced practice provider (APP). The tool helped facilitate a timely 6-month review for each provider by sending automated reminder emails, and it provided a central repository for current and historical reports, which made it easy to support audit activities such as Joint Commission surveys. The six new metrics related to copy-paste complemented an existing library of 27 inpatient and 26 ambulatory metrics used in the institution’s OPPE reports.

Results

From January 1-December 31, 2020, approximately 2.3M inpatient notes and 6.6M outpatient notes were authored in our EHR (Table 1). The notes were written by 4,103 providers. The average length of inpatient notes by character count was 5,194. Heart Institute notes were the shortest, on average, at 2,257 characters, while Medicine Institute notes were more than three times longer (6,876 characters on average). The average length of ambulatory notes was 2,869 characters. Cancer Institute notes were the longest (6,321 characters on average), and Diagnostic Medicine notes (2,084 characters) and Medicine Institute notes (2,214 characters) were the shortest.

Of the nearly 2.3M inpatient notes, 42% used copy-paste, and 19% of overall note content was copied. Of 6.6M ambulatory notes, 18% used copy-paste and 12% of note content was copied. Copy-paste use varied considerably by Institute/Specialty (ranging from 15% [Diagnostic Medicine] to 78% [Cancer] for inpatient notes and 11% [Medicine] to 59% [Cancer] for ambulatory notes).

The Institute with the highest usage of copy-paste in the inpatient setting was Cancer—of the 31,216 inpatient notes written by Cancer providers, 78% used some copy-paste, and 30% of note content was copied, on average. Cancer providers also used copy-paste frequently in the ambulatory setting (59% use, with 39% average copied content).

The Women and Children’s Institute (comprising OB/GYN, general pediatrics, and pediatric subspecialties) had the highest average % copied content (31%) for inpatient notes, but the lowest (7%) for ambulatory notes.

Within institutes, there was also considerable variation by provider in the use of copy-paste. Of the 4,103 providers who authored notes in the EHR from January-December 2020, 3,283 (80.0%) wrote at least 200 notes. Among the notes written by these 3,283 providers, use of copy-paste ranged from 0% to 100%. There were 571 providers (17.4%) who used copy-paste infrequently, in less than 5% of the notes they authored. On the other extreme, 333 providers (10.1%) used copy-paste in more than 75% of the notes they authored.

The large majority of providers who wrote at least 200 notes—2,406 of 3,283 (73.2%)—authored notes that included less than 25% total copied content (by character count). However, there were 88 providers for whom >50% of the content of their notes was inserted via copy-paste, and 8 providers for whom >75% of the content of their notes was inserted via copy-paste.
Table 1. Use of copy-paste by clinical institute. Nurse practitioners and physician assistants were not classified by clinical institute and are listed separately. CY=Calendar Year.

<table>
<thead>
<tr>
<th>Institute/Specialty</th>
<th>Inpatient Notes</th>
<th>Ambulatory Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total Notes in CY 2020</td>
<td>Notes with Copied Content</td>
</tr>
<tr>
<td>Anesthesia/Surgical Services</td>
<td>245,437</td>
<td>50,946</td>
</tr>
<tr>
<td>Cancer</td>
<td>31,126</td>
<td>24,297</td>
</tr>
<tr>
<td>Diagnostic Medicine</td>
<td>26,822</td>
<td>3,931</td>
</tr>
<tr>
<td>Heart</td>
<td>213,719</td>
<td>53,571</td>
</tr>
<tr>
<td>Medicine</td>
<td>872,776</td>
<td>433,143</td>
</tr>
<tr>
<td>Neuroscience</td>
<td>138,954</td>
<td>82,465</td>
</tr>
<tr>
<td>Orthopedics</td>
<td>83,192</td>
<td>26,267</td>
</tr>
<tr>
<td>Surgery</td>
<td>252,537</td>
<td>123,691</td>
</tr>
<tr>
<td>Women and Children's</td>
<td>171,741</td>
<td>79,662</td>
</tr>
<tr>
<td>Nurse Practitioner</td>
<td>58,222</td>
<td>25,222</td>
</tr>
<tr>
<td>Physician Assistant</td>
<td>189,145</td>
<td>55,031</td>
</tr>
<tr>
<td>Other</td>
<td>1,655</td>
<td>107</td>
</tr>
<tr>
<td>Grand Total</td>
<td>2,285,326</td>
<td>958,733</td>
</tr>
</tbody>
</table>

Figure 2 shows the inclusion of copy-paste metrics into our software tool for creating OPPE reports. The specific example in Figure 1 shows % Copied Content for Ambulatory notes. Use of copy-paste is shown over time. While we did not set targets for the six copy-paste metrics, the OPPE tool provided benchmarking information by comparing the provider’s use with that of his/her peers.

Figure 2. Example of how an individual provider’s use of copy-paste is included in the ongoing professional practice evaluation (OPPE; Conduent Care Management Inc., Tucson, AZ). The OPPE document containing this information is reviewed by the clinician’s supervisor (e.g., division chief) and discussed with the clinician every six months.

Discussion
Our analysis is one of the largest assessments of copy-paste utilization in the EHR, representing almost 9 million inpatient and ambulatory notes written by 4,103 physicians, APRNs, and PAs over one year across a broad clinical enterprise. Wang and colleagues analyzed 23,630 notes written by 460 clinicians in the inpatient medicine care setting at an urban academic medical center. They found that 46% of note context was copied, a much higher number than the 16% in our inpatient medicine notes. Wang also reported that direct care hospitalists wrote shorter notes (5,006 total characters) than medical students (7,053) and residents (6,720). We did not break out our note-writing data by
attending physician, resident physician, medical student, etc., but the average length of our inpatient medicine notes (6,876 characters) was on the high end of what Wang et al. reported.

In 2008, O’Donnell and colleagues surveyed 253 physicians who wrote inpatient notes electronically, finding that 90% reported using copy-paste, and 70% used it almost always or most of the time when writing daily progress notes. The vast majority of our providers used copy-paste at least occasionally, but only 42% of inpatient notes and 18% of ambulatory notes contained copied text. This difference is likely due to the improvements in EHR documentation capabilities, including templated text, that have been adopted since O’Donnell and colleagues’ survey in 2008.

**Use of Copy-Paste is a Poor Metric for Note Quality**

We were very careful in designing our OPPE reports to avoid framing copy-paste usage statistics as metrics for assessing provider competence or for inferring a connection to note quality. As Stetson and colleagues have argued, assessing quality of clinical documentation requires manual review. It’s possible, and probably common, for a short note to be “better” than a long note, and for a note that employed considerable copy-paste to be superior to a note where copy-paste was not used. Developing methods for assessing note quality is an important area for further research.

While we strongly caution against using targets associated with copy-paste behavior to evaluate clinician performance, we believe that identifying outliers (i.e., providers who seldom or never use copy-paste, and providers who use copy-paste far more often than their colleagues) may provide organizational leaders with targeted opportunities for further investigation. We believe that the OPPE process is an idea setting for such exploration and related conversations to take place. For example, if a provider rarely uses copy-paste, there may be an opportunity to educate her about the efficiency gains—and potential improvement in wellbeing—associated with appropriate use of the tool. Similarly, if a provider uses copy-paste far more extensively than his peers, a deeper dive may be warranted to assess whether the provider’s notes show evidence of misuse of the function.

**Appropriate Use of Copy-paste Can Reduce Documentation Burden**

The results of our study provide useful context for ongoing discussions among clinicians, informaticians, policymakers, and others about how to optimize the clinical documentation processes. In 2011, AMIA convened a policy meeting envisioning the future state of clinical data capture and documentation. The meeting produced a valuable set of guiding principles, including the admonition that “clinical data capture and documentation should be efficient and usable while enhancing the healthcare organization's and the care team's overall efficiency, effectiveness and productivity.” In January and February 2021, AMIA, the National Library of Medicine, Columbia University and Vanderbilt University hosted “25 By 5: Symposium to Reduce Documentation Burden on U.S. Clinicians by 75% by 2025,” with the stated objective to establish strategies and approaches to reduce the documentation burden on U.S. clinicians to 25% of the current level by 2025.

Holmgren and colleagues recently reported that the median U.S. clinician spends 90.2 minutes actively using the EHR per day—over 30 minutes more than non-U.S. clinicians. They concluded that U.S. clinicians’ greater EHR burden may be associated with nontechnical factors such as policies related to clinical documentation that may adversely impact clinician wellness. Promoting the appropriate and effective use of copy-paste and other EHR documentation support tools may help mitigate EHR burden and enable providers to focus on activities they perceive to be more impactful.

**Limitations**

Although we evaluated copy-paste practices for a larger number of providers than has previously been reported, they were all part of the same health system that had a single, well-established EHR. Our study did not assess differences in behavior among clinicians in different roles or stages of training (e.g., medical students vs. resident physicians vs. attending physicians). Additionally, the scope of our analysis was limited to note-writing and copy-paste use among physicians, PAs, and APRNs; many other members of the care team write clinical notes and may use copy-paste in different ways.
Conclusion

We developed a pragmatic approach to assess the use of copy-paste in aggregate, and provide individual training opportunities by including providers’ copy-paste usage statistics into the ongoing professional practice evaluation (OPPE) process that is required for hospital accreditation. This approach enables hospital administrators and other healthcare delivery organization leaders to monitor high-level usage of copy-paste at their institutions, and also provides individual feedback to providers about their copy-paste behavior relative to their peers. We believe that this process provides the appropriate balance in encouraging clinicians to use copy-paste as a time-saving tool while mitigating the associated risks.

References


10. CMS-1715-F: Revisions to Payment Policies under the Medicare Physician Fee Schedule, Quality Payment Program and Other Revisions to Part B for CY 2020, p. 380. Available at: https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/PhysicianFeeSched/PFS-Federal-Regulation-Notices-Items/CMS-1715-F


1184


Enhancing the IDEAS Framework with Ontology: Designing Digital Interventions for Improving Cancer Patients' Wellbeing

Nicole Veggiotti, MSc¹, Lucia Sacchi, PhD¹, Mor Peleg, PhD²
¹University of Pavia, Pavia, Italy
²Department of Information Systems, University of Haifa, Haifa, Israel

Abstract. Developing effective digital interventions to help patients form healthy habits is a challenging goal. IDEAS is a step-by-step framework that allows developers to draw ideas from intended users and behavioral theories, and ideate implementation strategies for them, followed by rapid prototype development. Based on our long experience with developing generic knowledge-based clinical decision support systems (CDSS) and integrating them with electronic health records (EHR) to deliver patient-specific advice, we observed a challenge that IDEAS is not addressing: the semantic detailing of the clinical knowledge behind the digital intervention and relevant patient data that could be used to personalize the digital intervention. To close the gap, we augmented two steps of IDEAS with an ontology that structures the target behavior as classes, derived from HL7 Fast Healthcare Interoperability Resources standard. We exemplify the augmented IDEAS with a case study taken from the Horizon 2020 CAPABLE project, that uses Fogg’s Tiny Habits behavioral model to improve the sleep of cancer patients via Tai Chi.

Keywords: Clinical decision support; Mobile Health; Knowledge Representation and Information Modeling; Controlled Terminologies Ontologies, and Vocabularies; User-centered Design Methods; behavior change; Fogg Behavioral Model

1. Introduction

People want to change their behaviors to improve their quality of life and make healthier choices. The most frequent behavior change areas [1] are smoking cessation, sleep quality improvement, weight loss, diet, physical activity, treatment adherence (i.e., adherence to medical recommendations including both taking medication and attending appointments), especially for chronic disease management. Digital interventions delivered through mobile technology can have great potential to facilitate such health behavior changes, as they provide improvements in efficacy, cost-effectiveness, safety, and scalability. Moreover, the literature highlights many successful digital interventions [2] developed for different types of behavioral changes such as Vegethon [3] for nutritional intervention, Happy Ending [4] for smoking cessation, SMART [5] for weight loss goal, and Headspace [6] for mindfulness. These systems cited above all utilize behavioral health techniques [7]. However, we have not encountered any ontologies that allow to detail the properties of the behavioral interventions that make them actionable in a general system.

To support the development of effective digital interventions for behavioral change, Mummah et al. [8] developed a framework and method called IDEAS (Integrate, DEsign, Assess, and Share). This framework uses a step-by-step approach to guide the development process using behavioral theory, design thinking, user-centered design, rigorous evaluation, and dissemination. While we find the IDEAS framework to be extremely constructive and useful, we believe it lacks formal and detailed structuring of digital interventions. In a previous study [9], we extended the IDEAS abstract concepts into concrete backend architectural components and graphical user-interface designs. In this paper, we propose an extension of IDEAS using an ontology to provide a conceptual framework for the definition of the behavior intervention. The behavioral theories exploited by the digital intervention and the clinical evidence for their success guide the structuring of the knowledge base. To structure the ontology, we rely on HL7 standards, and in particular on HL7's Fast Healthcare Interoperability Resources (FHIR) [10] standard. Defining a FHIR-based ontology allows standardizing the knowledge related to digital interventions for behavioural changes, which has not been done so far. Furthermore, this represents a step towards making the intervention actionable as an integrated part of a clinical decision support system (CDSS).

In this paper, we introduce the general methodology that we developed for the construction of such ontology, which is reusable in different application domains and for different target behaviors. We demonstrate our method through a case study taken from the Horizon 2020 CAncer PAatient Better Life Experience (CAPABLE) project (https://capable-project.eu/), where we are using Fogg’s Tiny Habits behavioral model [11] to improve the sleep of cancer patients.
2. Related work

The IDEAS framework

Mummah et al. [8] presented a ten-step guide to the development of effective digital health interventions that provides a disciplined way to incrementally translate behavioral theories into highly relevant and practical interventions. The ten steps are organized into a four-phases process as follows: Integrate phase, including (1) empathize with target users, (2) specify target behavior, (3) ground in behavioral theory; Design phase, including (4) ideate implementation strategies, (5) prototype potential products, (6) gather user feedback, (7) build a minimum viable product; Assess phase, including (8) pilot test to assess potential efficacy and usability, and (9) evaluation of the efficacy in an RCT; and Share phase, to (10) share intervention and findings.

Fogg’s Tiny Habits behavioral model

Fogg Behavioral Model (FBM) [11] is especially suitable to help patients turn interventions into habits. The model includes three main concepts: motivation, ability, and triggers. If all are present, they allow the behavioral change to occur and become a habit. Fig. 1 shows the core concepts of FBM. The vertical axis represents the motivation, which is driven by three different core motivators (pleasure/pain, hope/fear, and acceptance/rejection). The horizontal axis presents the ability of the subject to perform the behavioral task. Ability is affected by six different elements, called simplicity factors (time, money, physical effort, brain cycles, social deviance, and non-routine). The target behaviour has a trigger that prompts the subject to perform the behavioral task. Triggers can be a spark, a facilitator, or a signal.

This representation suggests, as the model asserts, that a person is able to achieve a target behavior if s/he has high motivation, high ability, and an effective trigger at the same instant; this places the person to the top-right of the curve, which is the actionable part where habits are formed. Note that in Fig. 1, three virtual interventions are shown as blue squares. Deep Breathing is not actionable as it is below the curve. In order to allow a person to adopt a behavior, we can increase her motivation or increase her ability by making the needed behavior smaller and easier to do.

Patient data models facilitating personalized decision support

In a previous CDSS that we developed, MobiGuide [12], we defined a patient model that included the patient’s clinical data, patient-reported symptoms, personal preferences, psychosocial context, and personal events. The model is both dynamic, as it reacts to current data from real-time monitoring of patient’s reporting and sensors tracking, and it is also highly adaptive to the personal preferences and contexts of the individual patient. The adaptive part of the model allowed us to suggest digital interventions that best fit with the patient’s characteristics. The model complied with HL7’s Virtual Medical Record.

Since that time, HL7 has released a more robust standard called Fast Healthcare Interoperability Resources (FHIR) [10]. It leverages existing logical and theoretical models to provide a consistent, easy to implement, and rigorous mechanism for exchanging data between healthcare applications. To achieve this goal, FHIR uses generic Resources (i.e., Patient, Observation), that define the properties of all exchangeable information.

![Fig.1. Fogg’s Behavioral Model with examples of the capsules for a specific subject, shown in blue. Adapted from [11].](image-url)
3. Methods

We augmented two steps of the IDEAS framework with an ontology structuring digital interventions for behavioral change. An ontology is an explicit and formal specification of a shared conceptualization representing a consensual and shared knowledge of an abstract model of a world's domain or phenomenon [13]. By defining classes in the ontology, we can specify the main concepts related to digital non-pharmacological evidence-based interventions, which we call "virtual capsules". In addition, we also represent a patient model useful for creating personalized interventions.

To define the structure of the classes in the ontology, we leverage the SNOMED-CT vocabulary and HL7 FHIR resources (a) Patient, (b) Medication_Request, (c) Medication, and (d) Dosage. The ontology provides a way to model a structured conceptual framework representing knowledge on specific concepts and the relations among them. This provides an instrument that allows more expressiveness that simply extending "isolated" FHIR resources. Moreover, using an ontology would allow extending and using the model with standards other than HL7 FHIR. Our framework is generalizable because it in principle allows representing any kind of non-pharmacological intervention. It is in fact possible to add other types of capsules targeting a wide variety of behaviours and goals.

To conform to the Basic Formal Ontology [14], Patient is an Independent_Continuant, Medication is a Process_Profile, Medication is an Object, and Dosage is a Generically_Dependent_Continuant. Medication_Request models an order for the medication and the instructions for administration of the medication to a patient, including dosage which also includes timing. Although, as observed in [15], FHIR does not address non-pharmacological behavioral interventions, we extend FHIR by using the analogy between pharmacological treatments and virtual Capsules representing behavioral interventions. These two types of interventions share several characteristics, except for the fact that pharmacological interventions use drugs, whereas the interventions in the capsules are lifestyle changes related to cognitive behavioral therapies or physical activity. Therefore, in our ontology, both Medication and virtual Capsule are modeled as subclasses of a new, more general class, called Treatment_Option (Fig 2b). Similarly, Treatment_Request (Fig 2a) is a new class that generalizes Medication_Request and the new analogous subclass Capsule_Request. As in FHIR, Treatment_Request includes dosage information. Fig 2c shows instances of Dosage for a Medication_Request and for a Capsule_Request.

Similar to the FHIR Medication resource, Treatment_Option has a codable concept that specifies the preferred name and controlled clinical vocabulary code (e.g., SNOMED-CT code) for the medications or the digital behavioral intervention (e.g., Mindfulness, Tai Chi), for which internal codes are used.

In addition, the Treatment_Option class structures the interventions by providing a goal-based and evidence-based view. The goal of Treatment Option was inspired by the Goal structure for CDSSs developed by Kogan et al. [16], which uses the relationship may_treat of the National Drug File Reference Terminology (NDFRT) [17]. Hence Treatment option has the property may_treat, whose domain is FHIR Observation (e.g., "may treat sleep problems" or "may treat fatigue"). For the Capsule subclass, goals are extended to wellbeing dimensions by adding the property fits_wellbeing, whose domain is Wellbeing_Dimension (i.e., mental, social, physical, spiritual, functioning, and global wellbeing [18]). The evidence-based view provides Treatment_Option with a property to cite evidence that the proposed intervention is effective for the goal (reference_effectiveness).

The Capsule class include additional properties corresponding to the theoretical behavioral model behind the intervention. The interventions that have been developed for the CAPABLE project are all based on the FBM [9]. To take this into account, the Capsule class has a set of properties that reflect factors related to trigger, ability, and motivation. The trigger (prompt) property can represent a signal, spark or facilitator. The two properties called effort_cognitive and effort_physical represent the FBM ability factors related to the mental and physical effort required for the specific behaviour. The two properties cost and in_out_doors are ability factors according to FBM. The FBM motivation factor is modeled through a reward function (has_reward) that assigns a score to the patient at the end of the execution of the practice.

Finally, the Patient class includes properties of the Patient FHIR resource, including name, gender, date of birth, marital status, care takers, and communication language. Similar to the patient model developed in [12], the Patient class also includes the main diagnosis and comorbidities or the patient, and the prescribed Treatment_Requests, which connects the patient to the patient-specific digital interventions and medications (Capsule_Request and Medication_Request). To allow goal-based reasoning [16] that is used in CAPABLE CDSS, we also add the clinical goals of the patient. As in [12], the Patient class also contains the observations related to the patient. Similar to the
Patient preferences used in [12], we represent the patient's ambiance preferences, musical preferences, technological preferences, and physical activity preferences, which allow personalization of recommendations.

Fig 2. The analogy between Medication_Request and (virtual) Capsule_Request for behavioral interventions. Our additions to FHIR resources are highlighted in yellow.

We developed the ontology via the Protégé knowledge modeling tool (protégé.standord.edu) version 3.4.8 using Web Ontology Language (OWL) version 1. Examples for instances of our extensions, for an intervention based on FBM [11], are presented in Section 4.

The presented ontology would fit well into two steps of the IDEAS framework (Fig. 3). The "specify target behavior" step of the "Integrate" phase is focused on making the target behavior highly specific and actionable, and providing the evidence for its potential impact. The goal of our ontology is to make the structure of the intervention explicit. In more detail, some of the properties of the Capsule class (may_treat and fits_wellbeing, for the goal-based view,
reference_effectiveness for the evidence_based view) and the timing property of the Dosage class are specifically devoted to indicate the target of the intervention and to ground its validity on scientific evidence.

The second step of the IDEAS framework that would benefit from the introduction of the presented ontology is the "ideate" step of the Design phase. This step is focused on the implementation strategies. To this end, the Capsule class is provided with a set of properties to explicitly refer to the behavioural model which is used to put the intervention into place (e.g., FBM). The Capsule_Request class is specifically aimed at modeling the implementation of the intervention on the specific user, which in turn modelled using the Patient class. The properties designed in the Patient class make it possible to propose and recommend the user the digital interventions that best fit with his intrinsic characteristics and preferences.

As shown in Fig. 3, the presented ontology extends the IDEAS framework in two specific steps. To get to the final structure of the ontology, though, we also slightly modified two other phases: Empathize and Gather. In IDEAS the users that are involved in collecting insights and requirements are patients, while we have relied on multiple stakeholders, including clinicians, psychologists, nutritionists, clinical and informatics researchers, and software developers. Each of them brought input to the requirement collection, as related to the virtual capsules (behavioral interventions).

Fig.3. The IDEAS framework [8], extended by our method. Changes are in blue.

4. Results in the CAPABLE Case Study

Most cancer patients are managed at home, facing long-term treatments that make the disease comparable to a chronic condition [19]. Patients are expected to assume a more significant role in managing their follow-up care. They must have the ability to control the symptoms and consequences of living with a chronic condition, including treatment, physical, social and lifestyle behavioral changes [20]. To improve patients' well-being [18] and to help them comply with behavioral interventions known to improve mental and physical health, we hypothesize that mobile technology can potentially support. Thus, we had ideated virtual capsules, which serve as the knowledge base of a CDSS patient coaching system.

These digital interventions were developed as part of a European project, part of the Horizon 2020 European program for research and innovation, called CAPABLE (CAncer PAtients Better Life Experience), that aims at building a CDSS to support cancer patients and their care providers during the home management of the disease, using Artificial Intelligence and Big Data potentialities.

In this paper we use CAPABLE as a case study for the design and the implementation of digital interventions using the ontology presented in Section 3. In our case study we profile a patient called, Maria Rossi, a hypothetical Italian female kidney cancer patient aged 66. Maria is modelled as an individual of the Patient class (Fig 4), and she has properties related to personal information (e.g., name, gender, marital_status, care takers, communication language) and life-style preferences (e.g., diet, physical activity, hobbies –following Wikipedia's list of hobbies [21], and technological preferences (i.e., PC/laptop, mobile phone, smart phone, tablet) (Fig 4A), clinical conditions (e.g., main
cancer condition and comorbidities of diarrhea and sleep problems) (Fig 4B), goals that the patient wants to achieve (Fig 4C), and prescribed treatments for these goals, including medications and virtual capsules (Fig 4D).

Fig. 4. Maria Rossi, an individual of the Patient class of the ontology

This profiling helps us to better associate users with personalized digital interventions. In fact, thanks to Maria’s preferences, it is possible to identify some capsules suitable for her, and others that might be inappropriate. As it is possible to see from the regular_physical_activity property value in Figure 4A, Maria does not practice regular physical activity. Tai Chi, which is a short physical and mental practice, could help her increase her level of physicality.
without too much effort. Moreover, since Maria likes to walk (physical_activity_preference property), a digital intervention such as the daily walk in nature can also fit well with Maria’s characteristics. Instead, a static capsule like Deep Breathing, a relaxation technique, is considered unsuitable for Maria, as it does not match her interests.

Fig. 1 shows how the capsules identified as suitable for Maria appear in the actionable part of the Fogg Behavioral Model, as the user has the necessary characteristics to have sufficient ability and motivation to carry out the interventions that are to the right of the curve. Capsules such as Deep Breathing appear below this actionable part, because Maria’s motivation for them is low; to make interventions more appealing to the patient it is necessary to modify them according to the user’s preferences and needs.

As previously mentioned, the Patient class has a property that connects it to individuals of the class Treatment_Request (has_treatment_request – Figure 4D). In the specific case of a behavioural intervention, this property allows us to link the Patient class with the Capsule_Request class, that models the knowledge related to digital intervention prescribed for a single patient.

One of the treatments that is requested for Maria is the Tai Chi (Fig 4D), which is modeled as an individual of the Capsule_Request class, as shown in Figure 2. In particular, Maria’s Capsule Request (Fig 2a) points to the Tai Chi Capsule (Fig 2b). Instances of the Capsule contain all the generic characteristics of the Tai Chi intervention, not strictly related to the user who has to practice it (Fig 2b). The dosage and timing properties of the recommended Treatment_Request (Fig 2a) are represented through an individual of the class Dosage (Fig 2c). The dose, rate, and timing can be used to personalize the intervention to the patient’s ability. For example, for some patients, 30 minutes of Tai Chi could be too much, but they could form a habit of practicing Tai Chi 5 minutes a day (dose) every day (rate) for one month, in the morning (timing).

The Tai-Chi Capsule instance for Maria is shown in Figure 5. It specifies the target behavior (Fig 5A) and the goals that the intervention aims to achieve (Fig 5B), through the properties may_treat and fits_wellbeing. These properties model that this capsule aims to improve sleep problems and fatigue thus improving physical and emotional well-being. Since every digital intervention developed in our work is based on scientific evidence, the Capsule class
contains the reference_effectiveness property (in Figure 5.C), which points at the benefits that the capsule can carry out to solve this issues (in this case the ameliorative effects that Tai Chi can bring in cancer patients).

Part D of Fig. 5 relates the Capsule to the FBM [11]. The trigger chosen for Tai Chi belongs to the Signal type, as notifications are sent to Maria, through the application, that remind her to perform the practice. The effort_cognitive and effort_physical related to the Tai Chi capsule are considered of medium intensity, as Tai Chi requires both mindfulness and body movement. The cost is low, as CAPABLE offers Maria some YouTube videos to perform the practice. Tai Chi may be practiced in- or out-doors. The FBM motivation factor is modeled through a reward function (has_reward) that assigns a score to Maria at the end of the execution of the practice, depending on the duration of the exercise.

Figure 6 presents the taxonomy of classes in our ontology and the properties of the Patient class, exemplified by the Patient individual of Figure 4.
4. Discussion

In this paper we presented the extension of the IDEAS framework with a structured ontology that allows developers of digital behavioral change interventions to address in a systematic and detailed way the specification of the behavioral change intervention (the Capsule), its dosage and timing (Treatment_Request), and the properties of the Patient that could be used to personalize the Capsule_Request to the patient’s characteristics and preferences. By basing our ontology on FHIR resources, we allow modelers and developers of digital interventions a standard way to specify CDSS knowledge in a way that is interoperable with electronic health records. We have used the modified IDEAS framework and the novel ontology that we had developed in our ongoing development of the CAPABLE prototype, which is now in its second iteration and will be used by patients and clinicians in 2023. Its preliminary evaluation has been done by over ten additional members of the CAPABLE project, including software developers, clinicians and patients.

The novel Capsule class structures the potential delivery model of the digital intervention according to Fogg Behavioral Model [1]. As part of the CAPABLE project we aim to use machine learning models that would learn how to dynamically personalize the customizable capsules by modulating the duration of the Capsule_Request or its intensity level in order to meet the patient’s "ability" factor of FBM. We have also started to develop methods for personalizing triggers by learning the user’s daily routine from sensors - when is she usually outside, when is she not moving enough, when is she sleeping, etc. This information could be used to predict the most suitable time for triggering (prompting) the reminder for the Capsule. We also started thinking about reward functions, as part of the motivation factor of FBM. Another part of the motivation is educating the patient about the benefits of adhering to the capsule recommendations (Capsule_Request).

To make the capsules more actionable, future work will target quantifying the effort required to complete the capsule (physical and cognitive) as well as the ability and motivation of the patient. The effort required for the different capsules can be determined in two ways. The first is the available evidence on the effectiveness of the activity in promoting behavioral change (e.g., we know that Tai Chi is effective if performed every day for 5 minutes, which implies a medium physical effort). The second way to determine the effort is to rely on the knowledge of the expert (physician, psychologist, physiotherapist) who is defining the capsule. This work lays the foundation for a CAPABLE’s virtual coaching system which will deliver personalized digital behavioral interventions to improve cancer patients’ wellbeing. Although our study is aimed at supporting the wellbeing of cancer patients, the behavioral change interventions that we offer could help all other chronic disease patients, whose physical health is often accompanied by mental distress. The mindfulness, nature-based, exercise-based, or positive psychology-based interventions that we offer are applicable to many other chronic disease patients. Furthermore, the work represents a step towards making digital interventions actionable as an integrated part of any knowledge-based CDSS. While we demonstrated our ideas using a particular case study, the ontology is generic and could be extended to support other behavioral models in addition to FBM.

Using the ontology would benefit developers and researchers of digital interventions beyond CAPABLE from the following aspects: (a) structuring their non-pharmacological intervention using a systematically-developed ontology leveraging on the available standards (FHIR, NDFRT, SNOMED-CT), (b) using the same ontology structure for different components of a complex CDSS in order to define the behavioural interventions but also to exchange information about its usage among system components, including the patient electronic health record, and (c) allowing a standard way for researchers to analyze data collected about the delivered interventions (both pharmacological and non-pharmacological) and their impact on the patient’s health and wellbeing in a unified way. Beyond the contribution of this research toward representation of digital behavioral health interventions via an ontological standard-based information model, the integration into the IDEAS framework supports the entire lifecycle of eliciting requirements till implementing a mobile health app.
Acknowledgement

The work described in this article has been funded by the European Union’s Horizon 2020 research and innovation programme under grant agreement No 875052 - CAPABLE (www.capable-project.eu).

References

Creating a Home for Genomic Data in the Electronic Health Record

Nephi Walton MD MS¹, Darren Johnson, MS², Bret Heale PhD¹, Thomas Person MS³, Marc Williams MD²; Intermountain Healthcare, Salt Lake City, UT 84107; Geisinger, Danville, PA 17822; Penn State University, State College, PA 16801

Introduction

Genomic data generated from massively-parallel sequencing have the potential to improve health through targeted clinical management. The integration of genetic information into healthcare is considered a major priority with several national and international initiatives focused on this goal[1]. Only recently have EHR vendors allowed for the storage and access of discrete genomic data. However, many challenges remain related to adoption and implementation of such technologies. Aside from storage of discrete genomic data there are new clinical workflows that are introduced by large scale population sequencing that have not been addressed by EHR vendors. Having worked through the integration of genomics for the two largest EHR platforms and for two large scale sequencing programs, the MyCode® Community Health Initiative (MyCode)[2] at Geisinger, and the Heredigene® Population Study (Heredigene)[3] at Intermountain Healthcare, we have performed an analysis of the current state and future needs for true genomic data integration.

Methods

At Geisinger through MyCode and in collaboration with the Electronic Medical Records and Genomics (eMERGE) network[4]we imported discrete pathogenic and likely pathogenic genomic variants for the Centers for Disease Control and Prevention (CDC) Tier one conditions (Hereditary Breast and Ovarian Cancer syndrome, Lynch syndrome, and Familial Hypercholesterolemia), and variants from seven pharmacogenes. The C282Y variant for hereditary hemochromatosis were also included to assess for challenges related to implementation of autosomal recessive conditions. We constructed context sensitive maintenance guidelines in the EHR for each of the disease conditions and built best practice alerts for medications impacted by the pharmacogenomic variants. Patient and provider facing information was developed for all disease variants as these were accessible through both the physician chart and the patient portal[5]. Similarly, at Intermountain Healthcare through Heredigene and the RxMatch pharmacogenomics platform we constructed a passive decision support system for 17 pharmacogenes and patient and provider facing information for CDC Tier one conditions and hereditary hemochromatosis. At both institutions, we designed data infrastructures and interfaces to handle the clinical workflow of the return of results process as neither EHR vendor had the capability to manage this workflow. Weekly calls were held with pertinent institutional stakeholders to address challenges and barriers to implementation as they arose. A list of these challenges was maintained throughout the process. After each implementation, we analyzed the noted challenges and identified barriers. Metrics were also gathered on usage of CDC for pharmacogenomic (PGx) at both institutions, and changes in management where tracking was possible.

Results

We were able to successfully implement genomics in both systems with compromises. We encountered vendor specific informatics challenges on both EHR vendor platforms including the inability to store discrete variants for pharmacogenomics, necessitating use of star allele diplotypes, and having inadequate infrastructure for autosomal recessive conditions when the patient is a compound heterozygote. Although these were vendor specific issues the vendors data implementations were based on legacy standards, which had deficiencies that contributed to these problems. Other global challenges include: 1) Need for a “home” for genomic data 2) Need for standardization of genetic phenotype. 3) Difficulty in sending discrete data from the genetic testing laboratory directly to the EHR. 4) Addressing the role of the Laboratory Information System (LIS) in the genomics process. 5) Inadequate and discordant standards for genomics in the HL7 genomic report format and the Fast Healthcare Interoperability Resource (FHIR) molecular sequence resource. 6) Challenges related to maintenance of patient/provider information and decision support in a rapidly changing field. 7) Lack of standard resources for patient/provider information. 8) Disparate implementation needs for clinical geneticists compared to primary care providers. 9) Different perception of discrete genomic data versus scanned PDF of the genetic report. 10) Difficulty in maintaining variant classification across patients and reports.
A technical infrastructure was designed to manage the return of results process at each institution that enabled 1) tracking of notifications of patients of their results, 2) genetic counseling processes, 3) provider notifications, 4) specialist referrals, 5) result based clinical actions, 6) cascade testing, 6) variant classification and reclassification.

Metrics on CDS usage in pharmacogenomics demonstrated that 73% (40/55) of providers changed prescriptions based on active decision support for new prescriptions but none of the providers changed dosing or medication when the patient had been established on the medication. Metrics for passive decision support demonstrated that usage was primarily restricted to those who ordered a PGx test for a specific purpose and only used after the initial test result came back and rarely used outside of the ordering provider or by the ordering provider for another indication.

**Discussion**

While we were able to achieve some success in integration of genomic data into the EHR, there are many challenges that need to be addressed. One of the biggest challenges moving forward is the long-term maintenance of gene/variant specific information and decision support. Genomic medicine is a rapidly changing field and EHR systems require support from vendor specific technical personnel to implement changes to patient/provider facing information and decision support. Engaging this technical team can be challenging given their responsibility to maintain all aspects of the EHR and they are generally not skilled in genomics requiring an alignment of teams to make any changes. EHR vendors adhering to more open standards for genomics and decision support, including SMART on FHIR, and CDS Hooks would allow for integration of external applications where maintenance can be performed by staff with genetics training and less technical skill. Another important component to this, is the need for maintained resources such as PharmGKB that are Infobutton compliant and can allow to quick access to the most recent recommendations and information on a specific gene or variant. To our knowledge no vendor has implemented Infobutton support for genomics, and outside of pharmacogenomics no such Infobutton compliant resources exist.

Difficulty in passing data from the laboratory to the EHR is a considerable challenge and one of the greatest components of this challenge is the standardization of the genetic phenotype (i.e., disease, metabolizer status, gene specific phenotype, modifier variant phenotype). We were able to capture discrete genomic data from reporting laboratories, however; laboratories used their own proprietary JSON structured format. Developing the ability to pass phenotypes required manually building a process for each phenotype and each lab, which is neither desirable nor sustainable as more laboratories and conditions are added. To effectively enable integration of genetic data in a scalable manner requires adequate standards for storing and transmitting genomic data and standard API’s that allow ancillary systems to interface with the EHR. While we have developed an external data architecture for managing the return of results process it is unclear at this time how or if this should be integrated into the EHR.

Active decision support was more effective in engaging providers and causing change, especially in situations of a new diagnosis or new medication. Passive decision support was less effective. We believe this was due to providers not seeking genomic information prior to prescribing even when it existed. Lack of training of providers about genomics likely contributed to this, as well as the relatively small number of patients that had genomic results making the information not common to seek. We believe as genomic information becomes more available and providers become better trained on its use that it will be sought after and must have an established place in the EHR similar to imaging or laboratory results. More research is needed to understand and optimize workflows for genomic medicine and to test novel methods of delivering genomic information to patients and providers.

**References**

Axes of Prognosis: Identifying Subtypes of COVID-19 Outcomes

Emma Whitfield, MMath1,2*, Claire Coffey, MPhil, BSc1,3*, Huayu Zhang, MSc4, Ting Shi, PhD4, Xiaodong Wu, MD, PhD5, Qiang Li, MD, PhD5, Honghan Wu, PhD1,2
1Health Data Research UK, London, United Kingdom
2Institute of Health Informatics, UCL, London, United Kingdom
3University of Cambridge, Cambridge, United Kingdom
4Usher Institute, University of Edinburgh, United Kingdom
5Shanghai East Hospital, Tongji University, Shanghai, China

Abstract

COVID-19 is a disease with vast impact, yet much remains unclear about patient outcomes. Most approaches to risk prediction of COVID-19 focus on binary or tertiary severity outcomes, despite the heterogeneity of the disease. In this work, we identify heterogeneous subtypes of COVID-19 outcomes by considering ‘axes’ of prognosis. We propose two innovative clustering approaches - ‘Layered Axes’ and ‘Prognosis Space’ – to apply on patients’ outcome data. We then show how these clusters can help predict a patient’s deterioration pathway on their hospital admission, using random forest classification. We illustrate this methodology on a cohort from Wuhan in early 2020. We discover interesting subgroups of poor prognosis, particularly within respiratory patients, and predict respiratory subgroup membership with high accuracy. This work could assist clinicians in identifying appropriate treatments at patients’ hospital admission. Moreover, our method could be used to explore subtypes of ‘long COVID’ and other diseases with heterogeneous outcomes.

Introduction

The clinical heterogeneity within COVID-19 patient outcomes has been demonstrated1,2, however, a great deal of this heterogeneity remains to be explored. Work has been done to classify patients and predict patient outcomes, much of which has been analysed by Wynants et al.3. However, this is often limited to predicting a binary or tertiary severity outcome (death/ICU/none). Our work furthers this by exploring more diverse poor prognosis outcomes.

For a variety of diseases, valuable insights can be gained by using unsupervised clustering to explore prognosis4,5. We propose and demonstrate a pipeline for exploring COVID-19 prognosis by using unsupervised clustering methods on patient outcomes to identify subtypes of poor prognosis. The identification of these subtypes facilitates the prediction of patient trajectories to nuanced poor prognoses. Our aim is thus to discover and predict a broader range of subtypes of COVID-19 prognosis, using a combination of supervised and unsupervised learning techniques.

Methods

Our methodology for this task consists of three steps: (1) Feature extraction from multimodal data: exploring, cleaning and manipulating our multimodal dataset in order to extract rich features; (2) Clustering on outcomes: using features collected at endpoints (discharge or death) to cluster patients, comparing results from a variety of clustering approaches using multiple axes of prognosis; (3) Classification at admission: using features at admission, mapping each patient to a cluster found in the previous step to predict their probable deterioration pathways. The pipeline of our work process is shown in Fig. 1. We will discuss each step in more detail below. Further details and code are available at https://github.com/knowlab/covid-subtypes.

Data

We demonstrate our methodology on a dataset consisting of 2815 health records of COVID-19 inpatients of Wuhan Sixth Hospital and Taikang Tongji Hospital, with admission dates between 4th February 2020 and 30th March 2020. It is worth noting that this cohort was treated in hospitals that were not overwhelmed, in a period with a treat-all policy in Wuhan, meaning that admission was routine for all COVID-19 patients. As a consequence, this cohort contains largely non-severe patients and the mortality rate of this cohort is 2.4%.

*These authors contributed equally to this work.
Pre-processing We removed duplicate records, leaving 2797 patients for whom a wide range of features were available, including general patient information, co-morbidities, smoking status (from free-text), symptoms (from free-text), lab test results, ICD-10 admission and discharge codes, and other prognosis features (death, ICU admission, supplementary oxygen, length of stay). Discharge ICD-10 codes were pre-processed and grouped into chapters. We removed codes ‘U07.1 - COVID-19’ and ‘Z22 - Carrier of infectious disease’ as, predictably, these were reported for the majority of patients. Discharge conditions were represented as one-hot vectors.

Data Imputation For the classification step, we wished to utilise much of the data available on patient admission, including laboratory test results and clinical measurements. However, we did not have readings for each measurement for each patient and so we imputed missing values. Our imputation technique was to randomly choose a value within the normal range for each test result, in order to reduce biasing the results. The values for normal ranges were found from various medical reference sources, the full list of which can be found at [https://github.com/knowlab/covid-subtypes](https://github.com/knowlab/covid-subtypes). We chose this, rather than using the average feature value, as we were advised by clinicians working at the hospitals that the absence of values is informative: a test is more likely to be performed if it is suspected the result will be outside the normal range.

Feature Extraction To prevent data leakage, we split features into two disjoint subsets: ‘outcome’ features for clustering and ‘admission’ features for classification. Outcome features took binary values, with the exception of length of stay, which was converted to two binary values: 00 for 0-2 weeks, 01 for 2-4 weeks, and 10 for 4 weeks+ (11 is not used). The admission features were taken from measurements in the first three days of hospital admission. Where available, the minimum and maximum readings for days 1 and 3 were used as separate features, otherwise, the average values from each day (1 and 3) were used.

Discovering COVID-19 Subtypes via Clustering

Axes of prognosis model We wished to cluster on features related to a patient’s outcome in order to explore nuanced subtypes of poor prognosis. To do this, we considered the idea that there are multiple ‘axes’ to prognosis. For example, the two ‘axes’ that make up an individual’s prognosis might be the system of deterioration (defined as clusters of diseases) and the severity of deterioration (quantified/qualified by proxy variables like mortality, types of treatments and length of stay). Here, if a patient had a respiratory discharge code, this would indicate the system of deterioration, whilst the use of oxygen therapy would be an indicator of the severity. To get a true picture of an individual’s prognosis, we must have a measure across all axes of prognosis - for example, ICD-10 chapters to indicate systems, and features such as ICU admission, death, oxygen therapies, and length of stay to measure severity. We will use different clustering approaches to leverage the idea of ‘axes of prognosis’.

The following example illustrates the potential value of this approach. Consider three patients who are admitted...
with COVID-19: patient A has no significant symptoms and is discharged after 5 days; patient B arrives with no significant symptoms, but is noted to be vitamin D deficient during their stay, recorded as ‘E55.9’, they are discharged after 7 days; patient C has already deteriorated significantly upon arrival and dies 1 day later. Clustering algorithms typically use distance measures to compare data points. If we consider these three patients in the space ‘Has Nutritional Code’ \(\times\) Death = \(\{0, 1\}^2\) then, for most distance measures \(d, d\text{(patient A, patient B)} = d\text{(patient A, patient C)}\). Thus, clustering algorithms lack context and can struggle to note the clearly significant distinction between outcomes. On the other hand, were we to consider the axes ‘system of deterioration’ and ‘severity of deterioration’ individually, or to transform the features into some space where distance can be ‘sensibly’ measured, a clustering algorithm could note the difference in severity of deterioration between patient C and patients A and B.

We propose the use of two novel methods using axes of prognosis: Layered Axes and Prognosis Space with the goal of exploring nuanced subtypes of poor prognosis. These are summarised below and in Fig. 2.

**Layered Axes:**

1. Choose axes of prognosis (e.g. system of deterioration, severity of deterioration, duration of illness) and assign each outcome feature to the axis of which it is most indicative (for example, ICD-10 chapters are indicative of the system of deterioration).
2. Choose one axis and, using only the features assigned to it, apply a standard clustering method to produce clusters. This produces clusters that clearly describe that axis.
3. Discard the features already used and choose a new axis. For each cluster already found, use the features assigned to the new axis and a standard clustering method to produce subclusters. These subclusters now clearly describe both axes.
4. The subclusters become the clusters and step 3 is repeated until all axes have been considered.

**Prognosis Space:**

1. Choose axes of prognosis and assign each outcome feature to the axis of which it is most indicative.
2. Define a ‘space transformation’ function to map features onto the axes - creating an interpretable Prognosis Space with fewer dimensions. This allows domain knowledge to be incorporated into the dimension reduction.
3. Use this function to map patients into the Prognosis Space.
4. Apply a clustering algorithm in the Prognosis Space. As we are no longer in a binary feature space, common distance measures, such as the Euclidean distance, have more meaning - this allows us to apply a wider range of clustering algorithms, such as DBSCAN\(^9\).

**Implementation** Our dataset consists of 2797 patients, for whom we have 16 binary features describing their outcome. As a baseline, we applied a standard clustering method to all features. Then, we applied both of our new methods to the Wuhan dataset and compared the clusters produced. Axes were chosen to demonstrate the potential different strengths of each approach. For the Layered Axes method, we chose our axes of prognosis to be system and severity of deterioration. We used ICD-10 chapter codes to indicate system of deterioration and ICU admission, death, use of noninvasive/invasive oxygen therapy and ECMO, and length of stay to indicate severity of deterioration. We clustered first on the system and then found subclusters using severity.

For the Prognosis Space method, we chose our axes of prognosis to be system of deterioration, need for oxygen therapy, severity of deterioration and duration of illness. As above, ICD-10 codes were used to indicate system; length of stay indicated duration of illness, ICU admission and death indicated severity of deterioration, and use of oxygen therapy and ECMO indicated need for oxygen therapy. We created a ‘space transformation’ function, \(f : \{0, 1\}^{16} \rightarrow \mathbb{R}^4\), that placed more weight on the severity and need for oxygen therapy axes. The function has the
basic form $f(x) = (f_1(x_{\text{system}}), f_2(x_{\text{OXY}}), f_3(x_{\text{severity}}), f_4(x_{\text{duration}}))$, where, for example, $x_{\text{system}}$ refers specifically to the features of $x$ that we used to indicate system. More details of the dimension reduction function used are given at https://github.com/knowlab/covid-subtypes.

As all our outcome features were binary and we wished to demonstrate the potential utility of this method in a clinical setting, we used the K-Modes clustering algorithm\(^{10}\) (implemented using the Python package kmodes\(^{11}\)) for our baseline and Layered Axes approach. This forms clusters based around a ‘centre’ - specifically the mode of the cluster - and uses the Hamming distance to compare points. The cluster centres are actual data points, which can then be used as a description for each cluster produced. The algorithm was run with several different numbers of clusters and the best result for each approach was chosen to reflect the heterogeneity in outcomes. Other clustering methods - such as K-Means clustering and DBSCAN - were also tested, however these algorithms are not optimised for the binary features used and so were less interpretable in our context.

Our Prognosis Space approach maps the binary features into $\mathbb{R}^4$ - this means it now makes sense to use the Euclidean distance to compare points, and a wider range of clustering algorithms become suitable. We demonstrated our Prognosis Space result using DBSCAN as it is more capable of capturing clusters of different shapes. Hyperparameters were optimised using the elbow method.

**Cluster Analysis** As this task is unsupervised, we had no knowledge of what ‘good’ clusters look like. Therefore, we sought clusters with a clear clinical interpretation, showing distinctions between different prognoses. Our cohort contained a large number of patients with mild symptoms: discounting length of stay, 1744 patients had 0s for all features. As such, for this cohort we expected at least one cluster with over 1000 patients, with others significantly smaller in size. To determine the clinical interest of a set of clusters, we examined heatmaps of feature prevalence for all the outcome features, alongside demographic feature prevalence.

**Predicting COVID-19 Subtype at Admission via Classification** For prediction, we used the clusters discovered as labels to generate a supervised classification problem to identify patient trajectories. We trained the classifiers using the results from our clustering methods, using cluster memberships as labels. We then input patient admission data to the classifiers, and cluster memberships were predicted by the classifiers, hence identifying likely patient outcomes. This not only provides information about potential patient deterioration pathways, which can be used to assist clinical decision making (e.g. how best to treat a patient), but also
provides a way to validate the clinical meaning of the clusters found.

Poor Prognosis Subtypes Since we were most interested in predicting the subtypes of poor prognosis, we focused only on the ‘interesting’ patients, which were decided using our understanding of the clinical meaning of the clusters: we disregarded individuals with no recorded discharge codes (since they had no prognosis subtypes on discharge), and performed multiclass classification on the remaining patients using their admission data to predict their deterioration pathway.

Classification Models For our system to be clinically useful, it is imperative the models are interpretable to clinicians. This motivated our choice of a set of ‘transparent models’ including decision trees, logistic regression and random forests\(^\text{12}\). We only present results on random forests as it performed the best. We used scikit-learn\(^\text{13}\) implementation. Random forests utilise many decision trees as an ensemble, which reduces potential issues with overfitting found in decision trees; they can provide better results than single decision trees since the variance between predictions is captured in the overall model. The space used and runtime is greater than that of a single decision tree, but this was not an issue for us due to the small dataset. The trees can be plotted, clearly showing decisions made by the classifier, including the importance of each feature used to aid in these decisions.

Classification Details We implemented grid search to choose the optimal hyperparameters for each classifier, and used scikit-learn\(^\text{13}\) stratified K-fold cross-validation to validate our models. The use of cross-validation was important due to the small cohort size: we wanted to make best use of the data by training and testing the model on multiple combinations of our data. The stratified variant of K-fold preserves the underlying distribution of the data, which we used due to the class imbalance observed. The value of \(k\) was set to 3; this was kept small to ensure there were enough samples in the test data for the least numerous classes in each fold, as some of the subtype classes had very few individuals.

Addressing Data Imbalance A limitation for the success of classification is the class imbalance that arises from many patients experiencing only mild symptoms. Since we are using real-world data, the imbalance of classes is representative of a real Chinese hospital cohort. However, to optimise the generalisability of our classifiers and increase performance, we used upsampling: we upsampled individuals in the least numerous classes in the training data in each fold by randomly sampling with replacement. Overall predictions are improved, however upsampling does not capture the variety of patients, potentially causing overfitting and not improving predictions for those deviating from the training set.

Results

Clustering

For the baseline and Layered Axes approaches, we used the K-Modes cluster centre descriptions for each cluster. For the Prognosis Space approach, we derived similar descriptions using the heatmaps produced. The descriptions of all the clusters found are given in Table 1, and heatmaps of the prevalence of binary features in each cluster are shown in Fig. 3.

Baseline Approach Most severe patients were put in cluster 2, meaning it failed to find subtypes of severe cases. There are two non-severe subtypes with no adverse prognoses: cluster 3 with 2-4 weeks of stay and cluster 1 with mixed stays, most within 2 weeks and a few 4+ weeks.

Layered Axes This method discovered several clinically sensible subtypes within groups with similar severity and background conditions. There are two clear subtypes of severe COVID patients: cluster 0b with high proportion of co-morbidities, particularly a combination of respiratory and nutritional; cluster 1a with mostly respiratory that were complicated with circulatory; the former was generally more severe. There are two subtypes of respiratory patients: cluster 1a - the severe group and cluster 1b - largely non-severe (40% discharged within 2 weeks). This is particularly interesting as respiratory conditions are widely recognised as a high risk-factor of severe COVID-19 cases despite the age group. There are also two subtypes of non-severe patients: cluster 3a with 2-4 weeks of hospital stay and cluster 3b with around 80% discharged in 2 weeks with no recorded complications.

Prognosis Space There are three very interesting subtypes of severe patients: cluster 5 - all stayed in ICU but all recovered; cluster 7 - quick deterioration to death within 2 weeks; cluster 8 - others all died after >2 weeks. For
**Table 1:** Clusters found using the three approaches - the standard baseline clustering approach, the Layered Axes approach and the Prognosis Space approach. Interesting subtypes are highlighted in bold text.

<table>
<thead>
<tr>
<th>Approach</th>
<th>Clustering Method</th>
<th>Cluster Label</th>
<th>Size</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>K-Modes</td>
<td>0</td>
<td>73</td>
<td>Digestive, los 2-4 weeks</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>1374</td>
<td>Non-severe</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>97</td>
<td>Respiratory, ICU, death, oxygen therapy</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>1018</td>
<td>Los 2-4 weeks</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>173</td>
<td>Circulatory</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>62</td>
<td>Digestive</td>
</tr>
<tr>
<td>Layered Axes</td>
<td>K-Modes</td>
<td>0a</td>
<td>150</td>
<td>Nutritional, non-severe</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0b</td>
<td>22</td>
<td>Nutritional, ICU, death, oxygen therapy</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0c</td>
<td>113</td>
<td>Nutritional, los 2-4 weeks</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1a</td>
<td>45</td>
<td>Respiratory, ICU, oxygen therapy</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1b</td>
<td>131</td>
<td>Respiratory, non-severe</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2a</td>
<td>90</td>
<td>Circulatory, los 2-4 weeks</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2b</td>
<td>120</td>
<td>Circulatory, non-severe</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3a</td>
<td>845</td>
<td>Non-severe, los 2-4 weeks</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3b</td>
<td>1281</td>
<td>Non-severe</td>
</tr>
<tr>
<td>Prognosis Space</td>
<td>DBSCAN</td>
<td>-1</td>
<td>37</td>
<td>Noise</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>914</td>
<td>Non-severe, los 2+ weeks</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>1114</td>
<td>Non-severe, los &lt;2 weeks</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>112</td>
<td>Oxygen therapy, los 2+ weeks</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>255</td>
<td>Circulatory, los &lt;2 weeks</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>232</td>
<td>Circulatory, los 2+ weeks</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>43</td>
<td>Respiratory, ICU, no death</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6</td>
<td>44</td>
<td>Oxygen therapy, los &lt;2 weeks</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7</td>
<td>28</td>
<td>Respiratory, death in &lt;2 weeks</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8</td>
<td>18</td>
<td>Many side effects, death in 2+ weeks</td>
</tr>
</tbody>
</table>

non-severe patients: cluster 1 patients have no underlying conditions and are discharged within 2 weeks; cluster 3 have mostly circulatory conditions (also many other conditions) and a speedy recovery (<2 weeks of stay); cluster 0 - others, with longer admission. There is another distinct finding - cluster 2 were those patients mostly on oxygen therapy but who recovered and never stayed in ICU.

**Classification**

We predicted poor prognosis subtypes of ‘interesting’ patients, disregarding patients without discharge codes or poor prognosis events. This contained classification sub-problems, based on the labels generated by Baseline K-Modes, Layered Axes K-Modes and Prognosis Space DBSCAN. For each of these, we classified using random forests and implemented grid search in order to find the optimal hyperparameters. These, along with additional experimental configurations can be found at our online resource: https://github.com/knowlab/covid-subtypes. A selection of our classification results are presented in Table 2, and Table 3 contains the most important feature for each classifier. Full classification results for all experiments and feature importances are also shown on our online resource.

**Layered Axes** The results for the Layered Axes clusters are shown in Fig. 4, with number of individuals classified into each cluster shown. These display an improved overall accuracy compared to baseline K-Modes, with an $F_1$ score of 0.627 on average. We built separate classifiers to predict the second-layer subcluster membership (within each top-layer cluster of nutritional, respiratory, and circulatory clusters). The $F_1$ score for the classification of respiratory patients’ subtypes is highest at $0.844$. In Fig 5, the confusion matrix for the classification of the most severe patients is shown; the subcluster sizes are very limited which impacts the robustness of these results.

**Prognosis Space** We present the results for the classification of severe patients’ subgroups using the Prognosis Space approach in Fig. 5. Subtypes 5 and 7 are classified with higher accuracy than 8, although these are the most severe patients. This is also highlighted in the greater $F_1$ score of 0.720 for distinguishing clusters 5 and 7, compared to that of 5, 7 and 8 (0.579). The cluster sizes are very limited which impacts the robustness of these results.
Figure 3: Heatmaps showing prevalence of binary features within each cluster found by each approach. OXY refers to use of oxygen therapy and LOS refers to the length of stay. Note that the features shown in the third row, Demographic information, were not used for clustering - they are shown only to aid interpretation of the clusters. For the Layered Axes K-Modes clusters, labels are in the form number-letter where the number indicates the clusters found from the first axis, and the letter the subclusters found within that cluster using features from the second axis. Note also that for the Prognosis Space approach, DBSCAN has been used and -1 indicated points described as ‘noise’.

Figure 4: Example confusion matrices heatmaps for Layered Axes K-Modes classification subclusters. Left: Nutritional; Middle: Respiratory; Right: Circulatory. x-axis: predicted label; y-axis: true label.

Discussion

Our approach is able to identify many subtypes of severe COVID-19 cases described with rich clinical features. We believe this is the first work that reveals the heterogeneity of COVID-19 with such detail. Many of these subtypes (such as clusters 1a and 1b for respiratory patients) were surfaced for the first time, conveying novel insights that lead to better understanding of the disease and could potentially guide us to treat patients with more personalised approaches. It is clear that the axes of prognosis approach plays a vital role in achieving this.
Table 2: Subtype classification results (macro average of multi-label results)

<table>
<thead>
<tr>
<th>Approach</th>
<th>Clustering method</th>
<th>Clusters</th>
<th>Acc.</th>
<th>Recall</th>
<th>Precision</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>K-Modes</td>
<td>0, 2, 4, 5</td>
<td>0.573</td>
<td>0.495</td>
<td>0.518</td>
<td>0.491</td>
</tr>
<tr>
<td>Layered Axes (all)</td>
<td>K-Modes</td>
<td>0, 1, 2</td>
<td>0.657</td>
<td>0.626</td>
<td>0.637</td>
<td>0.627</td>
</tr>
<tr>
<td>Layered Axes (severe)</td>
<td>K-Modes</td>
<td>0b, 1a</td>
<td>0.627</td>
<td>0.547</td>
<td>0.547</td>
<td>0.542</td>
</tr>
<tr>
<td>Layered Axes (respiratory)</td>
<td>K-Modes</td>
<td>1a, 1b</td>
<td>0.886</td>
<td>0.829</td>
<td>0.866</td>
<td>0.844</td>
</tr>
<tr>
<td>Prognosis Space (all)</td>
<td>DBSCAN</td>
<td>2, 3, 4, 5, 6, 7, 8</td>
<td>0.428</td>
<td>0.316</td>
<td>0.356</td>
<td>0.320</td>
</tr>
<tr>
<td>Prognosis Space (severe)</td>
<td>DBSCAN</td>
<td>5, 7</td>
<td>0.651</td>
<td>0.581</td>
<td>0.665</td>
<td>0.579</td>
</tr>
<tr>
<td>Prognosis Space (severe, respir.)</td>
<td>DBSCAN</td>
<td>5, 7</td>
<td>0.746</td>
<td>0.715</td>
<td>0.739</td>
<td>0.720</td>
</tr>
</tbody>
</table>

Table 3: Feature importances for the random forest classifiers. Impurity refers to the most important feature based on mean decrease in impurity, and permutation refers to the most important feature based on feature permutation.

<table>
<thead>
<tr>
<th>Approach</th>
<th>Most important feature (impurity)</th>
<th>Most important feature (permutation)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>Lactatedehydrogenase</td>
<td>Age</td>
</tr>
<tr>
<td>Layered Axes (all)</td>
<td>Systolic BP Day 3 Min</td>
<td>Cystatin C</td>
</tr>
<tr>
<td>Layered Axes (severe)</td>
<td>Serum sodium</td>
<td>Diastolic BP Day 3 Max</td>
</tr>
<tr>
<td>Layered Axes (respiratory)</td>
<td>Blood Sugar &amp; Oxygen Saturation</td>
<td>Blood Sugar D3 Min</td>
</tr>
<tr>
<td>Prognosis Space (all)</td>
<td>Lymphocyte %</td>
<td>Cystatin C</td>
</tr>
<tr>
<td>Prognosis Space (severe)</td>
<td>Age</td>
<td>Age</td>
</tr>
<tr>
<td>Prognosis Space (severe, respiratory)</td>
<td>Total Bilirubin</td>
<td>Total Bilirubin</td>
</tr>
</tbody>
</table>

Figure 5: Example confusion matrices for classification of severe patients. Left: Layered Axes K-Modes; Right: Prognosis Space DBSCAN. x-axis: predicted label; y-axis: true label.

Comparing to baseline K-Modes, using the Layered Axes approach can discover more nuanced clusters by grouping on one axis at a time and thus selecting a feature of significance from each axis. For example, at baseline, respiratory side effects are only considered in the context of very severe deterioration. In contrast, clusters 1a and 1b capture different outcomes for respiratory patients. It is interesting to note that, despite the very different outcomes of 1a and 1b, the proportion of over 65s is similar in both clusters. We might have expected that older patients were more likely to deteriorate significantly. Prognosis Space is also able to reveal more heterogeneity in outcomes. Clusters 5-8 seem particularly noteworthy for their similar yet distinct characteristics. In particular, clusters 7 and 8 both consist of patients who are admitted to ICU and die, but with quite distinct deterioration speed.

For classification, in Layered Axes, the classification results suggest that these nuanced multi-axis clusters not only make more clinical sense, but they also provide improved classification results in comparison to the baseline. In particular, the excellent classification results for the respiratory subclusters, 1a and 1b, provide a strong example use case. Here, we can predict accurately whether a respiratory patient is likely to deteriorate seriously leading to ICU and death (1a membership) or experience no severe prognosis (1b membership). This is especially interesting since...
there is no significant difference in the age of the patients between these two subclusters (Fig. 3). For Prognosis Space, results are promising, but lacking in robustness due to the larger number of clusters and therefore the smaller number of samples in each cluster. Clusters 5 and 7 both contain severe respiratory patients (5: ICU, 7: death) and can be distinguished with high accuracy by our classifier (Fig. 5) - but it struggles with cluster 8. Cluster 8 contains patients who deteriorate fatally, but not rapidly. Our classifier likely struggles with these individuals since they take a longer time to deteriorate, so their readings upon admission - which we use as classification features - may not be seriously irregular. The feature importances for each classifier also differ significantly. This is potentially clinically informative as it highlights certain test results which are informative for our classifiers (in most cases, even more so than clear risk factors such as age).

**Strengths** We have demonstrated that considering axes of prognosis allows for discovery of more nuanced subtypes of poor prognosis. In particular, our Layered Axes approach allows us to interpret every cluster found along each axis of prognosis being considered and is easily interpretable, especially when only a small number of axes are being used. On the other hand, our Prognosis Space approach allows us to incorporate domain knowledge and combine related binary features in a meaningful way. The Prognosis Space approach is comparable to other dimension reduction-clustering approaches such as Principal Component Analysis (PCA). We believe that compared with methods such as PCA, especially in the context of a binary feature space, our approach is both more interpretable and allows for the incorporation of targeted domain knowledge. Both of our approaches provide a more interpretable and clinically meaningful output than many ‘off-the-shelf’ clustering approaches.

**Limitations** We highlight the fact that the large majority of our cohort did not suffer from any significant side-effects from COVID-19. In fact, the clusters contain over 2300 patients who can be deemed to be ‘non-severe’ (clusters 1 and 3 in baseline K-Modes). This leaves only around 500 patients from whom we can derive more ‘interesting’ clusters and build classification models, leading to classifiers built from very small samples. Upsampling the smaller classes when training our models cannot capture diversity in test samples, so does not appear to be a good solution here. Therefore, our approach, especially when using the clusters found across multiple axes, must be tested on larger datasets.

Additionally, although we have extrapolated meaning from the clusters found, we cannot truly know which clustering is ‘best’, or even, ‘good’ since they are found in an unsupervised setting. If clustering is not optimal, our classification will likely also be worse - but this is a hard problem to overcome! A larger and more diverse dataset may also help with our confidence in clustering ability and predictions made. Additionally, there are potentially different endpoints and trajectories of deterioration that clustering approaches may be unable to pick up. Future work could explore other options for detecting these.

**Future Work**

This work provides a demonstration of our methodology for exploring heterogeneous disease prognosis. Therefore, potential future work is vast. Predominantly, these techniques need to be tested on a larger dataset, particularly with more patients with severe outcomes. This will likely improve the accuracy of clustering on severe patients, and the ability of classifiers to predict them. The methodology behind the Prognosis Space technique in particular should be applied to a larger dataset to properly explore its capabilities. Furthermore, a large dataset would allow for deep methods such as neural networks to be tested - these have been shown to perform very well on unsupervised learning problems with large datasets. Whilst we would not want to use these for our main predictions due to lack of interpretability, they could provide a useful comparison.

Secondly, our model may not be generalisable to cohorts from other countries, due to the different policies on hospital admission. However, we believe this approach could remain useful and provide interesting insights. Testing on different cohorts would enable exploration of the generalisability of the work. Further to this, our methodology needs to be tested and explored more rigorously. In different contexts, there are a vast number of clustering algorithms that could be used within our method - especially when continuous features are used. Moreover, the methodology could be refined and optimised for different contexts and diseases.

In some cases, classifiers could confidently predict patients who would die in a short time frame, yet struggled to identify those who would die after a longer period of deterioration, even if the outcomes were similar. Another aim for future work could be to establish whether using later time series points for these patients would allow correct
classification. We could also explore whether this would have potential applications for predicting long COVID. In particular, if a patient took a certain non-fatal prognosis pathway, this could be linked to the development of specific symptoms of long COVID later on.

**Conclusion**

We have proposed a novel *axes of prognosis* model and demonstrated how it could be used to identify diverse COVID-19 prognosis that accounts for the wide heterogeneity of the disease via a combination of clustering and classification. This methodology is not only suitable for predicting prognosis, but has a wide range of potential applications, for example, risk prediction and helping clinicians take preventative measures earlier on. This methodology could also be easily applied to a range of diseases with heterogeneous prognoses, including studying subtypes of long COVID.

**Acknowledgements**

CC and EW were supported by the Wellcome Trust (Grant Reference: 218529/Z/19/Z).

**Ethics Declaration**

This study was approved by the Research Ethics Committee of Shanghai Dongfang Hospital.

**References**


Design and development of an informatics-driven implementation research framework for primary care studies

Jiancheng Ye
Feinberg School of Medicine, Northwestern University, Chicago, USA

Abstract
The digitalization of the healthcare systems has resulted in a deluge of big data and has prompted the rapid growth of data science in medicine. Many informatics tools, such as data science, which is the field of study dedicated to the principled extraction of knowledge from complex data, can also introduce benefits into implementation science, quality improvement (QI), and primary care research. The increased amount of primary care QI initiatives, availability of practice facilitation-related data, the need for better evidence-based care, and the complexity of challenges make the use of data science techniques and data-driven research particularly appealing to primary care. Recent advances in the usability, applicability, and interpretability of data science models offer promising applications to implementation science. Despite the increasing number of studies and publications in the field, thus far there have been few examples of combining informatics and implementation framework to facilitate primary care studies. We designed and developed an informatics-driven implementation research framework to provide a coherent rationale and justification of the complex interrelationships among features, strategies, and outcomes. The proposed framework is a principle-guided tool designed to improve the specification, reproducibility, and testable causal pathways involved in implementation research projects in primary care settings.

INTRODUCTION
Implementation science
Implementation research is the scientific study of methods to promote the systematic uptake of research findings and other evidence-based practices into routine practice, and, hence, to improve the quality and effectiveness of health services and care. This relatively new field includes the study of influences on health care professionals and organizational behaviors. Implementation science shares many characteristics, and the rigorous approach, of clinical research. Implementation science in primary care involves investigation at multi-levels, and the targets of investigation include the patients, health care providers, primary care clinics, organization, community, society, and health system.

Informatics in primary care
Health care systems, especially primary care, suffer from major gaps between evidence and practice, unexplained practice variations, and suboptimal quality. Although health information technologies (HIT), such as information processing, and management are key to health care delivery and considerable evidence to improvements in health care quality and patient safety, primary care has a longstanding gap in the implementation of evidence-based practices and informatics packages. Informatics in primary care includes the development, installation, and implementation of electronic systems and relevant applications, including hardware, software, networking, and communication tools. In the past decade, IT development activities within the health care industry have increased as executives and providers recognized the urgent need for strategic information management and inadequacies of traditional information storage, retrieval, and analysis tools. However, most IT investments in primary care lack feasible guidelines for implementation and maintain sustainable improvement.

Implementation of informatics packages or tools in primary care has two main domains: (1) HIT infrastructure implementation. For example, implementing or updating electronic health record (EHR) systems to enable automated mechanisms for capturing data, and facilitate clinical decision making; (2) leveraging the clinical decision support (CDS) systems to improve health care providers’ performance and generate measurable metrics to evaluate the performance, such as efficiency, effectiveness, cost, efficacy, etc.

In general, most health care information systems are composed of automated billing and financial management, patient admission, discharge, transfers and registration, coordination of communications infrastructure, claims processing, customer service, and electronic data sharing. The informatics infrastructure needed for primary care includes feasible data collection methods, data repositories, clinical event monitoring, health care standards including standardized terminologies, digital sources of evidence, data-mining techniques, and communication technologies.
The informatics system in primary care needs to interface with, and assist, patients, health care providers, primary care managers, health care organizations, communities, and the public.

This study aims to design and develop an informatics-driven implementation research framework to provide a coherent rationale and justification of the complex interrelationships among features, strategies, and outcomes. We will adapt the typical evaluation logic model to integrate existing implementation science frameworks and informatics models as their core elements while keeping to the same aim of facilitating the causal modeling.

METHODS

Framework

To develop the informatics-driven implementation research framework, we adopted the Implementation Research Logic Model (IRLM) that was developed by Smith et al. This model enables a pipeline format for primary care systems that support the adoption and delivery of health practices by involving multiple levels factors within or outside the system and having its own unique characteristics. This framework also incorporated the key elements of the FITT (Fit between Individuals, Task and Technology) framework.7

Figure 1. Informatics-driven implementation research framework standard template
RESULTS

Features

Features are variables or factors that may impact the implementation (i.e., barriers and facilitators). All the features can be mapped onto the Consolidated Framework for Implementation Research (CFIR) framework. In some cases, features may act as moderators or mediators, thus indicating that they are links in a chain of causal mechanisms.

Informatics domains

Task

Tasks in primary care include collecting medical history, gathering preventive service information, completing forms, screening disease, diagnosing, tracking diagnostic data, making treatment protocols, proposing interventions, educating patients about self-care activities and medications, refilling prescriptions, receiving and resolving patients’ inquiries, referral tracking, and arranging home health care.

Technology

In primary care, technology and informatics system have several key elements. In general, a primary care clinic should have an available EHR to enable patients and providers to conduct shared decision-making. EHRs store personal health data, reliable patient-specific tools, and resources. EHRs provide every patient and their caregivers with the necessary information required for optimal care. They can help patients to better understand the complexity of medical care and more readily participate in clinical decision-making and preventive health behaviors. Electronic health information exchange (HIE) ensures security, privacy, and system compatibility. The exchange between organizations facilitates sharing of patient information at the point of the care delivery to eliminate unnecessary testing, improve safety, and facilitate efforts to improve health care quality and patient safety. Clinicians in primary care can leverage CDS can help health care providers utilize state-of-the-art medical knowledge in treatment decisions. CDS provides information management tools for the acquisition, manipulation, application, distribution, and display of appropriate patient- and task-specific clinical data to providers and patients that is conducive to correct, timely, and evidence-based clinical decision-making. Computerized provider order entry (CPOE) can help the tracking and analysis of health care processes. CPOE for tests, medicine, and procedures has the potential to decrease medical error, improve quality. It can help providers coordinate and collect patient-specific information. Population-based health care systems support the creation of large, integrated databases of patient-specific information that allow real-time management of populations of similar patients. These databases can facilitate the evaluation of new implementation strategies and provide insights into new associations between management approaches and health states.

All of these tools coordinate information dissemination and sharing from various databases to equip the provider in providing patient-specific, appropriate, timely, and evidence-based care.

Individual

For implementation research in primary care, there are several key stakeholders during the process, including primary care leaders, managers. Quality improvement practice facilitators are implementation scientists are also important because they work closely with primary care leaders and staff as partners and break down the research-practice divide in order to achieve the ultimate goal of increasing the public health impact. Table 1 demonstrates the informatics-driven interventions and strategies for implementation projects in primary care based on the three domains.
Table 1. Informatics-driven interventions and strategies for implementation projects in primary care

<table>
<thead>
<tr>
<th>Informatics domain</th>
<th>Interventions and Strategies</th>
</tr>
</thead>
</table>
| Task               | • Understanding, and incorporating, and improving the system’s workflow\(^\text{17}\)  
|                    | • Proactive strategies to report and solve problems in HIT applications in primary care  
|                    | • Achieving efficiencies in system use and evaluating the benefits  
|                    | • Determining routing of patient requests to appropriate providers within an integrated delivery system  
|                    | • Developing and updating patient care guidelines for prevention, diagnosis, and treatment to provide health care providers with essential information\(^\text{18}\)  
|                    | • Leaders in primary provide consistent support and feedback |
| Technology         | • Robust, secure, and available HIT infrastructures  
|                    | • Tools for patient-provider communication and consultation  
|                    | • Configuring the hardware and equipment to be conducive to performing manual tasks\(^\text{19}\) associated with the process  
|                    | • Feasible technologies to deliver benefits and support the health care delivery  
|                    | • Platforms for education and training\(^\text{20}\)  
|                    | • Combination of HIT and clinical operation  
|                    | • Automatic quality control  
|                    | • Capacities of protecting health information security and confidentiality |
| Individual         | • Collaborative relationships among health care providers, primary care leaders, HIT vendors\(^\text{21}\)  
|                    | • Organizational participants for major change  
|                    | • Environment for improving leadership and quality improvement  
|                    | • Continuously monitor and address users' concerns  
|                    | • Enhancing participation and perceived ownership of the system  
|                    | • Dedicated IT staff, and respected clinical staff  
|                    | • Culture of investing in change management |

Implementation strategies

**Intervention Characteristics**

Intervention Characteristics are key attributes of interventions that influence the success of implementation.\(^\text{22}\) Information technology management encompasses primary care, its infrastructure, strategies for use, and expected outcomes. Implementation research in primary care requires the support and resources of the organization and leadership. Most primary care systems have limited resources for QI. Resource allocation warrants close examination to maximize the benefits of the delivery of safe, effective, efficient, and high-quality care. Factors influencing organizational decisions surrounding implementation need to explore the characteristics of clinical conditions, practices, and settings, outstanding gaps, expected costs and benefits. The implementation research should be tailored to the needs of the organization. A balanced assessment of the effectiveness and costs of the system is needed to make successful implementation efforts.

**Inner Setting**

The inner setting is an active interacting facet and not just as a backdrop in implementation. Most primary care clinics are still at the stages ranging from considering adoption through early implementation. The lack of financial support for widespread informatics applications is considered a primary barrier to its implementation by both managers and clinicians.\(^\text{23}\) The financial burden of implementation, including acquisition and implementation costs, slow and uncertain financial payoffs, and disruption to clinical practices, is directly related to both the size of primary care and its readiness for conversion.

**Outer Setting**

An effective implementation needs stakeholders outside of the organization, such as policymakers, state and federal entities, and various participants (e.g., physicians, other providers, hospitals, payers, etc.).
Characteristics of Individuals

Organizations, including primary care, are made up of individuals. This framework can facilitate the understating of the interplay between individuals and their ripple effects through their teams, units, networks, and organizations on implementation. Clinicians desire a system that allows them to review and act upon test results safely and efficiently. Taking full advantage of informatics may require a team approach, where active involvement of interdisciplinary groups of providers and users is important. HIT use may empower patients in their exchange with providers and promotes the alignment of care between hospitals/clinics and patients' homes. Human factors such as system usability, process complexity, and user-engagement methods routinely influence the implementation.24

Process

A crucial first step for the organization in the implementation process is to assess both current informatics capabilities and estimate future needs.25 Equally important is the assessment of readiness for major organizational change, such as the ability to invest in change management and training, as well as the culture and processes needed to support implementation.26 Informatics systems to be implemented need to align with processes within the primary care to allow evolution and adaptation.27 On an individual level, the organization and its management need to involve end users' input in improving their work practices. They also need to consider other factors, such as the usability, usefulness, and flexibility of informatics tools and individualized training that will influence uptake. On a system level, a supportive culture, visibility of positive results from IT use, and a realistic timeline will enhance implementation.28

Outcomes

Outcomes include target clinical outcomes, health service outcomes, or indicators of implementation processes. Among these, implementation outcomes are the effects of deliberate and purposive actions to implement new treatments, practices, and services in primary care.29 The interrelated nature of implementation outcomes may present as non-linear, complex, or dynamic sequences of adoption by a delivery agent. Outcomes earlier in the sequence can be conceptualized as mediators and mechanisms of strategies on later implementation outcomes. Specifying which strategies are theoretically intended to affect which outcomes, through which mechanisms of action, is crucial for improving the reproducibility and generalizability of implementation research in primary care and validating the results.

DISCUSSION

A number of exciting challenges face the rapidly developing field of health and biomedical informatics in primary care. Integration of implementation science methodologies into primary care will speed the development of evidence-based interventions that have demonstrable public health impacts. Testing and validating the theories that underlie implementation efforts is needed to enhance the development of next-generation analytic methods and interventions. This informatics-driven framework used systems thinking that allows us better understand how each element in the system interacts and impacts each other, thus proposing intervention strategies accordingly.

The proposed informatics-driven implementation research framework provides a compact visual depiction of implementation research studies. Its use in the planning, executing, reporting, and synthesizing of implementation research could increase the rigor and transparency of complex studies that ultimately could improve reproducibility—a challenge in the field—by offering a common structure to increase consistency and a method for more clearly specifying links and pathways to test theories.6 The proposed framework would be a useful and generalizable guideline for future practice facilitation projects, QI initiatives, and health care intervention implementation studies.

CONCLUSION

We designed and developed an informatics-driven implementation research framework to provide a coherent rationale and justification of the complex interrelationships among features, strategies, and outcomes. The promise of implementation science lies in the ability to conduct rigorous and reproducible research, to clearly understand the findings, and to synthesize findings from which generalizable conclusions can be drawn and actionable.
recommendations for practice change emerge. The proposed framework is a principle-guided tool designed to improve the specification, reproducibility, and testable causal pathways involved in implementation research projects in primary care settings.

References

Characterizing Brain Signals for Epileptic Pre-ictal Signal Classification

Hao Yu¹, Shize Jiang, MD², Yan Huang, PhD³, Xiaojin Li, PhD³, Xiaoling Wang, PhD¹, Liang Chen, MD PhD², Jin Chen, PhD³
¹Shanghai Key Lab of Trustworthy Computing, East China Normal University, Shanghai, China ²Department of Neurosurgery, Huashan Hospital, Fudan University, Shanghai, China ³Department of Neurology, University of Texas Health Science Center at Houston, Houston, TX ⁴Institute for Biomedical Informatics, University of Kentucky, Lexington, KY

Abstract  Epilepsy is a kind of neurological disorder characterized by recurrent epileptic seizures. While it is crucial to characterize pre-ictal brain electrical activities, the problem to this day still remains computationally challenging. Using brain signal acquisition and advances in deep learning technology, we aim to classify pre-ictal signals and characterize the brain waveforms of patients with epilepsy during the pre-ictal period. We develop a novel machine learning model called Pre-ictal Signal Classification (PiSC) for pre-ictal signal classification and for identifying brain waveform patterns critical for seizure onset early detection. In PiSC, a unique preprocessing procedure is developed to convert the stereo-electroencephalography (sEEG) signals to data blocks ready for pre-ictal signal classification. Also, a novel deep learning framework is developed to integrate deep neural networks and meta-learning to effectively mitigate patient-to-patient variances as well as fine-tuning a trained classification model for new patients. The unique network architecture ensures model stability and generalization in sEEG data modeling. The experimental results on a real-world patient dataset show that PiSC improved the accuracy and F1 score by 10% compared with the existing models. Two types of sEEG patterns were discovered to be associated with seizure development in nocturnal epileptic patients.

1 Introduction

Epilepsy, as one of the most common neurological diseases, can be characterized by a combination of complexity and intractability. It is a persistent or secondary change in the brain and can cause serious consequences in neurobiology, psychology, sociology, and cognitive dysfunction. Improving the diagnosis of epilepsy and providing better treatment has been a significant and challenging research problem around the world. Electroencephalogram (EEG) is one of the main means to observe the electrophysiological activity of the brain, although, for different purposes, the way of obtaining brain signals could vary. Common forms of brain signals include scalp electroencephalogram (EEG), electrocorticography (ECoG), and Stereo-electroencephalography (sEEG). Among them, sEEG is a minimally invasive surgical procedure used to identify areas of the brain where epileptic seizures originate. During sEEG, doctors place electrodes in targeted brain areas, which are then monitored to precisely locate the source of the seizure. sEEG can find seizure sites deep in the brain that a regular EEG test may not reach.

In recent years, with the rapid development of deep learning technologies, deep neural networks have been extensively used in EEG analysis. However, there is still little work on sEEG. One of the most important reasons is that sEEG data contains strong individual characteristics, which may interfere with the judgment of the model. Furthermore, in the existing studies, the pre-ictal and the inter-ictal periods of epilepsy have not been completely unified. Although a trained neurologist can recognize seizure-specific signals 30-60 minutes before the seizure, the size of the time window depends on the sensitivity of the algorithm and signal significance. Finally, compared to EEG databases with hundreds of patients, sEEG data is scarce. As far as model training is concerned, it is more challenging to train a deep learning model on small data while maintaining its generalizability.

Leveraging recent advances in brain signal analysis, we propose a deep learning algorithm called Pre-ictal Signal Classification (PiSC) that can distinguish seizures from non-seizure using sEEG signals a certain time before seizure onset (see Figure 1(b)). The goal is to identify sEEG patterns closely related to pre-ictal signals. Knowing the seizure patterns in sEEG will help doctors improve the location of epileptic areas in patients and understand the underlying mechanism of seizure initiation and spread. As shown in Fig 2, PiSC has two phases: a deep learning model to classify the pre-ictal signals and a pattern recognition procedure to identify and categorize critical sEEG patterns related to seizure onset.
Figure 1: sEEG data acquisition and segmentation. (a) an example of the raw sEEG data where the x-axis is time, the y-axis is channel name, and the dotted line indicates seizure onset, and (b) sEEG data segmentation where data for both pre-ictal signals (red) and no-seizure (blue) are separated into 2-second segments. All the consecutive 2-second segments in a 95-second sEEG sliding window will be used for voting.

In summary, our main contributions are described below:

- We propose PiSC, a deep learning framework that integrates data harmonization and meta-learning. While the former effectively mitigates patient-to-patient variances, the latter can fine-tune a trained epileptic pre-ictal signal classification model using a few cases from a new patient. The design improves the model training stability and model generalizability.

- A unique preprocessing procedure is developed to convert raw sEEG signals to data blocks ready to apply deep learning and traditional machine learning models for seizure pre-ictal signal classification. PiSC can automatically identify pre-ictal brain signal patterns in the raw sEEG data that could be critical for the development of seizure prediction or seizure detection tools.

- Experimental results show that PiSC has increased the accuracy and F1 on seizure classification by at least 10% compared with baselines.

- Two types of novel sEEG waveforms that correlate well with seizure development have been identified.

2 Related work

sEEG directly captures electrical signals on the brain through electrodes implanted in the deep part of brain tissue. Compared with scalp EEG, sEEG can provide richer information and has a higher signal-to-noise ratio. The procedure usually involves implanting multiple electrodes into the patient’s brain, and there are usually 8-16 contacts along the electrode axis, with a center-to-center distance of 3.5mm.

To automatically characterize seizure patterns from brain signals, brain signal processing tools have been developed. However, most of these works are focused on scalp EEG data, which cannot be directly adapted to analyzing sEEG data. This is mainly because of the technical differences on data acquisition. EEG channel positions are common for almost all patients. Whereas the implantation of sEEG electrodes is based on the anato-electro-clinical analysis and the potential epileptic area of each patient. Needless to say, sEEG has a higher sampling rate than EEG.

Currently, only a few studies have studied seizure patterns based on sEEG data. Among them, EDFbrowser, MNE, and Brainstorm are multi-platform universal browsing applications with rich and intuitive graphical interfaces and convenient toolkits for the analysis of time series sEEG or EEG data. These platforms support neurophysiological data exploration, visualization, and analysis. Sharma et al. is an sEEG data analysis model that predicts seizure timing and identifies epileptic areas. Zhang et al. proposed a model that uses the power spectrum of to classify the sEEG. To our knowledge, the recently developed deep learning models have not been adapted to advance sEEG data characterization.
Finally, we build a set of fixed-length sEEG segments $X$ where $L$ are channels sorted according to the order in brain, $C$ channel in $C$ white matter cortex, which are two major brain functional areas. Mathematically, let $X_{win}(c_i, r)$ be a window function to convert a long sEEG signal into multiple fixed-size data blocks with natural orders on both rows and columns, facilitating the application of deep learning models on sEEG data.

Next, we sort all the sEEG channels in the gray (or white) matter by constructing a minimal spanning tree $T^g$ (or $T^w$) where each node is a channel and the channel-channel distance is measured using physical coordinates in the Talairach coordinate space. By traversing through $T^g$ (or $T^w$) with the depth-first search, we sort all the channels in $T^g$ (or $T^w$), saved in channel list $L^g$ (or $L^w$) so that roughly speaking, channels that are close to each other in the Talairach coordinate space are also close to each other in $L^g$ (or $L^w$). The output is sEEG signal matrix $Z$, where rows are channels sorted according to the order in $L^g$ and $L^w$, and columns are time points.

Finally, we build a set of fixed-length sEEG segments $X$:

$$X = win(fl t(sam pl(Z, r), f_{low}, f_{high}), \omega)$$

where $sam pl(.)$ is a resampling function with sampling rate $r$; $fl t(.)$ is a band passing function where $f_{low}$ and $f_{high}$ are low and high frequency thresholds; $win(.)$ is a window function to convert a long sEEG signal into multiple

### 3 Method

To classify epileptic pre-ictal signals and to identify seizure patterns, PiSC has two phases: to classify pre-ictal events, and to identify and categorize critical patterns in epileptic pre-ictal signals. The model architecture is shown in Fig. 2.

#### sEEG Data Preprocessing

A data preprocessing procedure is developed to rearrange the raw sEEG data based on the functional area of the brain and the channel location in the Talairach coordinate space, such that long and continuous raw sEEG signals can be transferred into multiple fixed-size data blocks with natural orders on both rows and columns, facilitating the application of deep learning models on sEEG data.

First, we separate all the sEEG channels (see example in Fig. 1(a)) based on whether they are in the gray matter or in white matter cortex, which are two major brain functional areas. Mathematically, let $C^g$ be a set of $n$ channels located in the gray matter cortex of the brain, $C^g = \{c_1^g, c_2^g, c_3^g \ldots c_n^g\}$, where $c_i^g$ represents the physical coordinates of the $i^{th}$ channel in $C^g (i = 1\ldots n)$. Similarly, let $C^w$ represents the set of $m$ channels located in the white matter cortex of the brain, $C^w = \{c_1^w, c_2^w, c_3^w \ldots c_m^w\}$, where $c_j^w$ represents the physical coordinates of the $j^{th}$ element in $C^w (j = 1\ldots m)$.

Next, we sort all the sEEG channels in the gray (or white) matter by constructing a minimal spanning tree $T^g$ (or $T^w$) where each node is a channel and the channel-channel distance is measured using physical coordinates in the Talairach coordinate space. By traversing through $T^g$ (or $T^w$) with the depth-first search, we sort all the channels in $T^g$ (or $T^w$), saved in channel list $L^g$ (or $L^w$) so that roughly speaking, channels that are close to each other in the Talairach coordinate space are also close to each other in $L^g$ (or $L^w$). The output is sEEG signal matrix $Z$, where rows are channels sorted according to the order in $L^g$ and $L^w$, and columns are time points.

Finally, we build a set of fixed-length sEEG segments $X$:

$$X = win(fl t(sam pl(Z, r), f_{low}, f_{high}), \omega)$$
PiSC Phase 1. Pre-ictal Signal Classification

PiSC Phase 1 consists of two data harmonization models and a classification model. While the former are encoder-decoder designed to mitigate patient-to-patient variances in the training data, the latter is a CNN enhanced with few-shot learning to improve the learning effectiveness on new patients using only a few samples. The model architecture is shown in Fig. 2(a).

Using a dedicated encoder-decoder structure for the positive training data (i.e. pre-ictal) and the other for the negative training data (i.e. no-seizure), we expect to preserve critical features that describe the differences between positive and negative while minimizing patient-to-patient variances. Let sEEG segment \( x_i \) \((x_i \in X)\) be the input and \( z \) be the hidden representation with parameter \( \theta \), the encoder can be denoted as \( q_\theta(z|x) \). Using the hidden representation \( z \) as the input, the decoder \( p_\phi(\tilde{x}|z) \) reconstructs the probability distribution of sEEG segments, where \( \phi \) is the parameter of the decoder. The models are trained to minimize the loss \( L_{ed} \):

\[
L_{ed} = \frac{1}{N} \sum_{i=1}^{N} L_i(\theta, \phi)
\]  

(2)

where \( N \) is the total number of sEEG segments and

\[
L_i(\theta, \phi) = -E_{z \sim q_\theta(z|x_i)} [\log p_\phi(\tilde{x}_i|z)] + KL(q_\theta(z|x_i) \parallel p(z))
\]  

(3)

where the first part is the reconstruction loss, and the second part is a regularizer. While the patient-to-patient variances in training can be reduced using two encoder-decoders, the testing data need to feed directly to the classifier. To maintain the same data dimension for training and testing, we follow the same strategy in Rusu et al. 21 that the decoder output \( \tilde{x} \), instead of the hidden representation \( z \), is used as the encoder-decoder output. Fig. 4(a,b) shows an example of the sEEG segment \( x \) and its corresponding decoder output \( \tilde{x} \).

The second component of PiSC is meta-learning. To train a classifier (e.g. CNN) using limited training data and then adapting it for new patients, we define a two-way \( K \)-shot problem with the episodic formulation in MAML. 19. Specifically, PiSC is trained iteratively with all the tasks \( T_{mtn} \), where \( T_{mtn} \) in \( T_{mtn} \) represents a task instance that includes \( K \) positive and \( K \) negative cases drawn randomly from the training data (see Algorithm 1). In the model testing stage, we randomly split every new patient’s sEEG signals into small support set \( S \) and a large query set \( Q \).
Testing task instance $T^m_{i\text{ttn}}$ consists of $K$ positive and $K$ negative cases in $S$ and $2K$ testing cases in $Q$. The trained PiSC is then fine-tuned with all the tasks $T^m_{i\text{ttn}}$ (see Algorithm 2).

The loss function of PiSC is defined as:

$$L_{\text{total}} = \lambda_\beta (L^p_{\text{ed}} + L^n_{\text{ed}}) + \lambda_\alpha L_T$$

(4)

where $L^p_{\text{ed}}$ is the loss of the negative encoder-decoder trained with no-seizure (negative) data; $L^n_{\text{ed}}$ is the loss of the positive encoder-decoder trained by pre-ictal (positive) data; $\lambda_\alpha$ and $\lambda_\beta$ are hyperparameters; and $L_T$ is the cross entropy loss of PiSC defined as:

$$L_T = -\frac{1}{N}\sum_{i}^{N} [y_i \ln(F(\tilde{x}_i)) + (1 - y_i) \ln(1 - F(\tilde{x}_i))]$$

(5)

where $\tilde{x}$ represents the input of CNN; $y$ is the label of $\tilde{x}$; and $F(\tilde{x})$ returns the predictive value of $\tilde{x}$ using PiSC.

Finally, majority voting is used to obtain the result of every sEEG sliding window. An sEEG sliding window consists of many non-overlap fixed-size consecutive sEEG segments. For every sEEG segment, its label can be predicted using PiSC. For all the segments in the same sliding window, we compare the number of positives and negatives. If more than a certain percentage of the segments are positive, the whole sliding window is labeled as pre-ictal.

**PiSC Phase 2. sEEG Pattern Identification**

To identify critical seizure pre-ictal patterns in the raw sEEG data, we develop a three-step procedure shown in Fig. 2(b). In the first step, for every positive sEEG segment that is correctly classified, we generate a localization map and identify hot spots using the Gradient-weighted Class Activate Mapping (Grad-CAM) algorithm22. In Grad-CAM, the gradients flow into the final convolutional layer to produce a coarse localization map highlighting important regions in the input data (see Fig. 4(c)). In the localization map, each hot spot indicates the channels and time points critical for pre-ictal signal classification. In the second step, we map each hot spot in a localization map to the raw sEEG data using both the temporal range and the channel information, resulting in raw sEEG fragments critical for pre-ictal signal classification. Fig. 4(d) shows an example of localized raw sEEG signals. Finally, we align all the hot spots using dynamic time warping (DTW) and separate them into multiple clusters using hierarchical clustering. The optimal number of clusters can be determined using Davies-Bouldin Index (DBI)12. The representative sEEG segment of each cluster can be identified using the scoring function in Eq. 6:

$$\text{Cluster Score}(x_i) = 1 - \text{ReLU}(DTW(x_i, DBA(X)))$$

(6)

where $X$ is a set of the sEEG segments in a cluster, $x_i$ represents the $i^{th}$ segment in $X$, $DTW(\cdot)$ aligns and compares any two waveforms using dynamic programming23, $DBA(\cdot)$ identifies the gravity center of all the sEEG segments in $X^{24}$, and $ReLU(\cdot)$ is the linear activation function25. Segment with the highest score is selected as the cluster center.

### 4 Experimental Results

**Data and Model Implementation**

sEEG data were collected from 42 epileptic patients in the Huashan Hospital affiliated to Fudan University, Shanghai, China. The data include 20.4 hours of no-seizure signals and 5.8 hours of pre-ictal signals. Among all the patients,
Table 2: The performance of PiSC and all the methods-to-compare on pre-ictal signal classification using sEEG data with relatively long duration (~one hour).

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F1 score</th>
<th>ROC AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Matching Nets</td>
<td>0.5148±0.0270</td>
<td>0.5110±0.0400</td>
<td>0.5841±0.0305</td>
<td>0.5442±0.0235</td>
<td>0.5182±0.0217</td>
</tr>
<tr>
<td>Relation Nets</td>
<td>0.5960±0.0048</td>
<td>0.6071±0.0108</td>
<td>0.5893±0.0122</td>
<td>0.5942±0.0083</td>
<td>0.6928±0.0082</td>
</tr>
<tr>
<td>Prototypical Nets</td>
<td>0.6631±0.0208</td>
<td>0.6796±0.0222</td>
<td>0.6695±0.0369</td>
<td>0.6704±0.0218</td>
<td>0.7426±0.0359</td>
</tr>
<tr>
<td>MAML</td>
<td>0.7276±0.0152</td>
<td>0.7506±0.0124</td>
<td>0.7297±0.0174</td>
<td>0.7392±0.0122</td>
<td>0.7560±0.0113</td>
</tr>
<tr>
<td>PiSC w/o MAML</td>
<td>0.5514±0.0024</td>
<td>0.6102±0.0026</td>
<td>0.4326±0.0355</td>
<td>0.4993±0.0023</td>
<td>0.5258±0.0019</td>
</tr>
<tr>
<td>PiSC w/o vote</td>
<td>0.7605±0.0061</td>
<td>0.7822±0.0095</td>
<td>0.7589±0.0116</td>
<td>0.7704±0.0065</td>
<td>0.7983±0.0122</td>
</tr>
<tr>
<td>PiSC</td>
<td>0.8256±0.0074</td>
<td>0.8259±0.0152</td>
<td>0.8266±0.0141</td>
<td>0.8258±0.0091</td>
<td>0.8260±0.0086</td>
</tr>
</tbody>
</table>

Table 3: The performance of PiSC and all the methods-to-compare on pre-ictal signal classification using sEEG data with relatively short duration (~15 minutes).

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F1 score</th>
<th>ROC AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Matching Nets</td>
<td>0.5123±0.0168</td>
<td>0.5095±0.0384</td>
<td>0.5366±0.0318</td>
<td>0.5212±0.0296</td>
<td>0.5247±0.0174</td>
</tr>
<tr>
<td>Relation Nets</td>
<td>0.6832±0.0071</td>
<td>0.7511±0.0120</td>
<td>0.6862±0.0154</td>
<td>0.7170±0.0100</td>
<td>0.8127±0.0064</td>
</tr>
<tr>
<td>Prototypical Nets</td>
<td>0.6996±0.0192</td>
<td>0.7259±0.0174</td>
<td>0.7072±0.0249</td>
<td>0.7144±0.0163</td>
<td>0.7594±0.0308</td>
</tr>
<tr>
<td>MAML</td>
<td>0.7558±0.0079</td>
<td>0.7809±0.0148</td>
<td>0.7629±0.0163</td>
<td>0.7687±0.0129</td>
<td>0.7991±0.0072</td>
</tr>
<tr>
<td>PiSC w/o MAML</td>
<td>0.5547±0.0060</td>
<td>0.5354±0.0102</td>
<td>0.5435±0.0146</td>
<td>0.5330±0.0327</td>
<td>0.6083±0.0038</td>
</tr>
<tr>
<td>PiSC w/o vote</td>
<td>0.8077±0.0047</td>
<td>0.8386±0.0109</td>
<td>0.8117±0.0080</td>
<td>0.8247±0.0047</td>
<td>0.8133±0.0102</td>
</tr>
<tr>
<td>PiSC</td>
<td>0.8499±0.0097</td>
<td>0.8692±0.0098</td>
<td>0.8371±0.0109</td>
<td>0.8526±0.0083</td>
<td>0.8497±0.0076</td>
</tr>
</tbody>
</table>

Figure 4: sEEG segment and its localization map. (a) sEEG segment $x$. (b) encoder-decoder output $\hat{x}$. (c) localization map where hot spots (red) indicate the channels and time points critical for seizure onset classification. All were visualized with Python matplotlib with the default color scheme. (d) raw sEEG signal in the hot spot indicated channel.

six nocturnal epileptic patients contributed 85% of pre-ictal data and 100% no-seizure data (see patient information in Table 1), while the other epileptic patients contributed 15% pre-ictal data. The no-seizure signals were collected when patients were sleeping or resting and did not experience any seizure. The pre-ictal signals were collected at most one hour before seizure onset. The data within 30 seconds before seizure onset were discarded (see Fig 1(b)).

PiSC was built on Pytorch and deployed on a server with Intel i7 CPU and Nvidia GeForce RTX 1080 GPU. It took an average of 1.5 hours to complete the model training. The code and manual of PiSC are available at https://github.com/jinchen74/SEEG_LSTM.
Figure 5: Trend of accuracy of pre-ictal signal classification. sEEG data were aligned to the last time point. The 0 at the x-axis was 30 seconds before seizure onset. It shows that on long sEEG data (P1 and P4) PiSC’s performance was above 90% at one hour before seizure onset, and then dropped sharply to about 70% at 10 minutes before seizure onset, indicating that there may be a turning point before seizure onset where the sEEG patterns may change. The results on short sEEG data (P5 and P6) show the performance of PiSC is high and relatively stable, indicating that the local abnormal discharge patterns, once captured by the model, can be used to infer seizure onset precisely.

Figure 6: sEEG segment clustering. The raw sEEG segments critical for the classification of pre-ictal signals were separated into five clusters. The visualization of the cluster centers reveals that there were two types of brain discharge patterns i.e., spikes in cluster 1 and 3 and high frequency oscillations in cluster 0, 2, and 4.

In data preprocessing, the length of sEEG segments was two seconds, and the length of sEEG sliding windows for voting was 95 seconds. sEEG re-sampling rate was 100 Hz, and band passing thresholds were 0 Hz and 30 Hz respectively. The outputs were 130-by-200 data blocks. In PiSC phase 1, the hyperparameters were $\lambda_\alpha = 0.8$, $\lambda_\beta = 0.2$, $\alpha = 0.001$, $\beta = 0.01$, and $\gamma = 0.001$. To avoid model overfitting, a simple CNN was selected as the classifier. The CNN includes four convolutional layers, four max-pooling layers, four ReLu regularization functions, and a dropout (0.5) layer. In PiSC phase 2, the optimal number of clusters identified using DBI$^{12}$ is five. A representative sEEG segment closest to each cluster center was used for further analysis.

Pre-ictal Signal Classification Results

In practice, a trained model shall be applied to new patients for seizure onset classification. To evaluate the performance of PiSC on new patients, we used the leave-one-out approach. Precisely, all six nocturnal epileptic patients but one were used for training, and the last nocturnal epileptic patient and all the non-nocturnal epileptic patients were used for model testing. In testing, the ratio of positive to negative samples is 1. We fed every 95-second sEEG sliding window of the testing patient, which included multiple short and consecutive sEEG segments, to a trained PiSC. By aggregating PiSC’s outcomes for all the short segments using voting, we predicted the label for that sliding window and compared it with the ground truth. After assessing all the sEEG sliding windows, we computed precision, recall,
the model performance decreased sharply from above i.e. 30 seconds before seizure onset. For P1 (red) and P4 (green), whose sEEG data duration was about one hour, testing data collected before or after the changing point. Here, all the sEEG data were aligned to the last time point, mechanisms that trigger epileptic seizures. As shown in Fig. 5, the performance of PiSC varied dramatically on the average, PiSC increased by 10% on accuracy, indicating that in the PiSC architecture, the data harmonization model precisely identified pre-ictal patterns and classified pre-ictal signals precisely one hour before seizure onset. On short sEEG data respectively. It shows that although there is no standard definition of the duration of pre-ictal period, PiSC precisely identified pre-ictal patterns and classified pre-ictal signals precisely one hour before seizure onset. On average, PiSC increased by 10% on accuracy, indicating that in the PiSC architecture, the data harmonization model (Fig 2(a) left) mitigated patient-wise differences, meta-learning (Fig 2(a) right) fine-tuned the trained PiSC for new patients, and voting on a long sliding window improved model performance. In addition, we found that channel shorting plays an important role on signal convolution. Without channel sorting, the accuracy of the PiSC w/o MAML model dropped to 45% on P1, which is 10% lower than the same data with channels sorted. The results also show that there could be a changing point around 10 minutes before seizure onset. It may be associated with the underlying mechanisms that trigger epileptic seizures. As shown in Fig. 5, the performance of PiSC varied dramatically on the testing data collected before or after the changing point. Here, all the sEEG data were aligned to the last time point, i.e. 30 seconds before seizure onset. For P1 (red) and P4 (green), whose sEEG data duration was about one hour, the model performance decreased sharply from above 85% to about 70% at roughly 10 minutes before seizure onset, suggesting that the sEEG patterns before and after the changing point were distinctively different from each other. For P5 (purple) and P6 (brown) with short sEEG duration (~15 minutes), the performance of PiSC was high and relatively stable, showing that the local abnormal discharge patterns, once captured by the model, might be used to infer seizure precisely.

**Pre-ictal Pattern Identification in Raw sEEG**

In PiSC phase 2, we separated all the positive sEEG segments correctly classified by PiSC into five clusters. The representative segments of each cluster are shown in Fig. 6. Two neurosurgeons specialized in epilepsy independently examined 50 sEEG segments mostly close to the cluster centers (see the pair-wise similarity in Fig. 7(a)). It was found that there were two types of sEEG waveforms in the data. The first type, as shown in clusters 1 and 3, is a type of interictal epileptiform discharges that usually indicates a potential area with seizure activity nearby. The second type was thought to be high-frequency oscillations (HFOs), a potential electrophysiological biomarker to improve focal epileptic brain mapping, as shown in clusters 0, 2, and 4. Surgical removal of regions that generate HFOs is more likely to be seizure-free. These nocturnal seizure patterns in raw sEEG may reveal regions with a high probability

**Figure 7: Distribution of pre-ictal signal clusters in relatively long (~1 hour) and short (~10 minutes) sEEG data.** (a) Heatmap shows the pre-ictal signal sEEG segment clustering result. (b) Cluster 3 was abundant when the patient was away from seizure and then it decreased gradually when seizure was approaching. On the contrary, the proportion of cluster 4 increased with time. (c) Cluster 4 quickly accumulated when seizure is about to occur. The 0 at the x-axis is 30 seconds before seizure onset.

accuracy, F1, and ROC AUC. We repeated the process until all the patients were used as testing. We compared PiSC with four existing few-shot learning algorithms, i.e. Matching Network (using advances in attention and memory to achieve rapid learning), Prototypical Network (learning a metric space to compute distances between prototype representations of each class), Relation Network (adopting few-shot learning for image classification and object detection), and MAML (model-agnostic meta-learning) given that a very small amount of training data were provided due to the nature of sEEG. Refer to the original papers for their detailed network architectures.

Table 2 and 3 summarize the performance of PiSC, two ablation models, and four compared models on long and short sEEG data respectively. It shows that although there is no standard definition of the duration of pre-ictal period, PiSC precisely identified pre-ictal patterns and classified pre-ictal signals precisely one hour before seizure onset. On average, PiSC increased by 10% on accuracy, indicating that in the PiSC architecture, the data harmonization model (Fig 2(a) left) mitigated patient-wise differences, meta-learning (Fig 2(a) right) fine-tuned the trained PiSC for new patients, and voting on a long sliding window improved model performance. In addition, we found that channel shorting plays an important role on signal convolution. Without channel sorting, the accuracy of the PiSC w/o MAML model dropped to 45% on P1, which is 10% lower than the same data with channels sorted. The results also show that there could be a changing point around 10 minutes before seizure onset. It may be associated with the underlying mechanisms that trigger epileptic seizures. As shown in Fig. 5, the performance of PiSC varied dramatically on the testing data collected before or after the changing point. Here, all the sEEG data were aligned to the last time point, i.e. 30 seconds before seizure onset. For P1 (red) and P4 (green), whose sEEG data duration was about one hour, the model performance decreased sharply from above 85% to about 70% at roughly 10 minutes before seizure onset, suggesting that the sEEG patterns before and after the changing point were distinctively different from each other. For P5 (purple) and P6 (brown) with short sEEG duration (~15 minutes), the performance of PiSC was high and relatively stable, showing that the local abnormal discharge patterns, once captured by the model, might be used to infer seizure precisely.

**Pre-ictal Pattern Identification in Raw sEEG**

In PiSC phase 2, we separated all the positive sEEG segments correctly classified by PiSC into five clusters. The representative segments of each cluster are shown in Fig. 6. Two neurosurgeons specialized in epilepsy independently examined 50 sEEG segments mostly close to the cluster centers (see the pair-wise similarity in Fig. 7(a)). It was found that there were two types of sEEG waveforms in the data. The first type, as shown in clusters 1 and 3, is a type of interictal epileptiform discharges that usually indicates a potential area with seizure activity nearby. The second type was thought to be high-frequency oscillations (HFOs), a potential electrophysiological biomarker to improve focal epileptic brain mapping, as shown in clusters 0, 2, and 4. Surgical removal of regions that generate HFOs is more likely to be seizure-free. These nocturnal seizure patterns in raw sEEG may reveal regions with a high probability
of seizure generation and could be further applied in a clinical setting.

Further analysis reveals that one of the spikes and one of the HFOs were strongly correlated with seizure development. Fig. 7 shows the temporal distribution of the five pre-ictal sEEG segment clusters in the short and long sEEG data. In Fig. 7b, cluster 3 (yellow), a spike pattern, was abundant when the patient was about 1 hour to 40 minutes before seizure onset and then it gradually decreased when the seizure onset was approaching. On the contrary, the proportion of cluster 4 (blue), an HFO pattern, increased with time. Both patterns correlated well with the trend of PiSC accuracy shown in Fig. 5. Moreover, when a seizure was going to occur (within 10 minutes, see Fig. 7c), the proportion of cluster 4 increased quickly. The results suggest that cluster 3 could be special spikes indicating the patient is a few hours away from seizure and cluster 4 HFO might be a local abnormal discharge pattern that quickly accumulated only when seizure is about to occur. These computer identified new brain wave patterns may be critical for the development of seizure prediction or seizure detection tools.

5 Conclusion

Epilepsy is a kind of neurological disorder characterized by recurrent epileptic seizures. Studies have attempted to recognize brain signal patterns related to seizure using brain signal data. However, reliably prediction or early detection of seizure onset remains a challenging problem. Leveraging high-quality sEEG data and recent advances in deep learning, we develop a novel deep learning algorithm called PiSC. In PiSC, we convert raw sEEG signals to data blocks ready for seizure onset classification, develop a novel deep learning architecture to classify the event of seizure using sEEG data, and identify critical seizure patterns in the raw sEEG data. The architecture of PiSC consists of two data harmonization models to learn the patient-to-patient variances in training and a meta-learning framework to enhance the effectiveness of model fine-tuning. The experimental results show that PiSC is significantly better than the existing models, and the sEEG waveforms identified by PiSC may potentially advance our understanding of the underlying mechanisms of the occurrence of a wide range of epileptic seizures and for guiding anti-epileptic therapy.

PiSC could be further improved in the following aspects. First, the temporal relationships among sEEG segments (critical for seizure classification) have not been considered. We will model segments with varying lengths and model their temporal relationships in the next version of PiSC. Second, sEEG segment embedding may unnecessarily increase the model complexity. We wonder if the data embedding problem can be addressed using a simpler approach. Also, PiSC could be further evaluated by comparing it with the existing EGG-based pre-ictal classification models.

For that, we will extend PiSC to fit EEG or ECoG data and further test the model using public seizure prediction data, including the Freiburg seizure prediction EEG database and the Kaggle seizure prediction database.

References


A Study of Social and Behavioral Determinants of Health in Lung Cancer Patients Using Transformers-based Natural Language Processing Models

Zehao Yu, Xi Yang, Chong Dang, Songzi Wu, Prakash Adekkanattu, Jyotishman Pathak, Thomas J. George, William R. Hogan, Yi Guo, Jiang Bian, Yonghui Wu

1Department of Health Outcomes and Biomedical Informatics, 2Division of Hematology & Oncology, Department of Medicine, College of Medicine, 3Cancer Informatics Shared Resources, University of Florida Health Cancer Center, University of Florida, Gainesville, Florida, USA; 4Information Technologies and Services, 5Department of Population Health Sciences, Weill Cornell Medicine, New York, NY, USA.

Abstract

Social and behavioral determinants of health (SBDoH) have important roles in shaping people’s health. In clinical research studies, especially comparative effectiveness studies, failure to adjust for SBDoH factors will potentially cause confounding issues and misclassification errors in either statistical analyses and machine learning-based models. However, there are limited studies to examine SBDoH factors in clinical outcomes due to the lack of structured SBDoH information in current electronic health record (EHR) systems, while much of the SBDoH information is documented in clinical narratives. Natural language processing (NLP) is thus the key technology to extract such information from unstructured clinical text. However, there is not a mature clinical NLP system focusing on SBDoH. In this study, we examined two state-of-the-art transformer-based NLP models, including BERT and RoBERTa, to extract SBDoH concepts from clinical narratives, applied the best performing model to extract SBDoH concepts on a lung cancer screening patient cohort, and examined the difference of SBDoH information between NLP extracted results and structured EHRs (SBDoH information captured in standard vocabularies such as the International Classification of Diseases codes). The experimental results show that the BERT-based NLP model achieved the best strict/lenient F1-score of 0.8791 and 0.8999, respectively. The comparison between NLP extracted SBDoH information and structured EHRs in the lung cancer patient cohort of 864 patients with 161,933 various types of clinical notes showed that much more detailed information about smoking, education, and employment were only captured in clinical narratives and that it is necessary to use both clinical narratives and structured EHRs to construct a more complete picture of patients’ SBDoH factors.

Introduction

People’s health outcomes are associated with complex, multi-level factors including clinical, biological, social, and health behaviors. To measure health outcomes, transdisciplinary approaches that can leverage a wide range of potential factors are needed. Major initiatives such as the Patient-Centered Outcomes Research Institute (PCORI)-funded PCORnet1 and the Observational Health Data Sciences and Informatics (OHDSI) have been established to promote research using electronic health records (EHRs) for health outcomes research. Clinical factors (e.g., diagnoses, medications) captured in EHRs have been widely used in various clinical studies. Due to lack of resources, patient’s biological, social and health behaviors are under-studied in health outcomes-related studies, which may cause potentially confounding issues (in statistical analyses) or misclassification errors (in machine-learning based classifiers). The Electronic Medical Records and Genomics (eMERGE) network and the NIH All of Us Research program have been established to develop resources of genetic data linked with EHRs. On the other side, through well-established conceptual frameworks such as the social-ecological model4 and the NIMHD Minority Health and Health Disparities Research Framework5, patient’s social determinants of health (SDoH; e.g., education, employment, income disparities) and their health behaviors (or behavioral determinants of health – BDoH; e.g., smoking, alcohol use) are increasingly recognized as important factors influencing health outcomes.6 There is an increasing interest to examine social and behavioral determinants of health (SBDoH) in shaping people’s health.

SBDoH are important risk factors affecting people’s health and healthcare outcomes. The most consistent predictors for the likelihood of death in any given year is level of education7 and poverty, estimated to account for 6% of US mortality8. In cancer, the second leading cause of death in US, up to 75% of cancers occurrences are associated with SBDoH factors9. Studies have reported that many SBDoH contribute to individual cancer risk, influence the likelihood
of survival, and affect cancer early prevention and health equity. A recent study reported that SDoH factors such as poverty, lack of education, neighborhood disadvantage, and social isolation play important roles in breast cancer stage and survival. Many SDoH factors are also associated with the screening of cervical cancer, breast cancer, and lung cancer. Reports from US institutes (e.g., Institute of Medicine, HealthyPeople 2020, and Health and Human Services) and international organizations (e.g., World Health Organization) reflect the increasing consensus of acknowledging SBDHo as significant contributory factors. The World Health Organization (WHO) defined structured codes to capture some of the SBDHo in EHRs. One potential source of SBDHo in the EHRs is in the International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM) Z codes (Z55–Z65).

In February 2018, ICD-10-CM Official Guidelines for Coding and Reporting approved that all clinicians, not just the physicians, involved in the care of a patient can document SBDHo using these Z codes. In our previous study, we have examined the use of ICD10 Z codes in the OneFlorida Clinical Research Consortium and reported a low rate of utilization for these Z codes (270.61 per 100,000 at the encounter level and 2.03% at the patient level). In order to advance cancer control and prevention, a transdisciplinary approach that integrates biological, clinical, social, and behavioral factors is needed. Nonetheless, SBDHo factors are scarcely and inconsistently documented in structured EHRs despite EHR systems provided opportunities for them to be manually entered as discrete data, but are often available in clinical text. Natural language processing (NLP) systems that systematically extract SBDHo factors from unstructured clinical text are needed to better assess cancer outcomes.

NLP is the key technology to unlock this critical information embedded in clinical narratives to support various downstream clinical studies that depend on structured data. NLP has received great attention in recent years. Various clinical NLP systems, such as MedLEE (Medical Language Extraction and Encoding System), MetaMap, KnowledgeMap, and cTAKES (Clinical Text Analysis and Knowledge Extraction System), have been developed to extract medical concepts from clinical narratives. To ensure the accuracy of information extraction, researchers have invested great effort into the development of clinical NLP methods. Clinical NLP systems approach the extraction of medical concepts from clinical narratives as a Named-Entity Recognition (NER) task. NER first identifies the boundaries (start position and end position in text) of medical concepts and then determines their semantic categories (e.g., diseases, medications). NLP methods based on statistical machine learning (ML) models have been increasingly applied and demonstrated good performance. Recently, NLP methods based on deep learning (DL) models have demonstrated superior performance than traditional ML models. However, most of the NLP systems focused on clinical factors (medical concepts directly generated by clinical practice, e.g., diseases and medications); NLP methods to extract SBDHo factors have been under-studied.

There are limited studies exploring NLP methods to extract SBDHo from clinical narratives. Feller et al. developed machine-learning classifiers to determine whether 11 categories of SBDHo are presented in clinical documents. Steer et al. also compared 5 machine-learning classifiers in classifying whether or not 5 categories of SBDHo presented in clinical text. Recently, Lybarger et al. developed an SBDHo corpus that consisted of 4,480 social history sections and applied deep learning-based NLP methods to extract 12 categories of SBDHo. To the best of our knowledge, no study has been reported investigating the difference between SBDHo extracted from clinical narratives and those captured in structured EHRs. This study aims to explore a state-of-the-art NLP model, clinical transformers, to extract SBDHo from clinical narratives. Using cancer as a study case, we (1) systematically examined two state-of-the-art transformer-based NLP models in extraction SBDHo from clinical narratives, (2) applied a transformer-based NLP model to extract SBDHo concepts for a lung cancer screening patient cohort, and (3) examined the difference between SBDHo extracted by NLP and those captured in structured EHRs. This is one of the earliest studies to apply transformer models to extract SBDHo and examine the difference between clinical narratives and structured EHRs.

Methods

Dataset

This study used EHR data from the University of Florida (UF) Health Integrated Data Repository (IDR) from 1999 to 2020, including both structured data and clinical narratives. Supported by the UF Clinical and Translational Institute (CTSI) and the UF Health, the UF Health IDR is a clinical data warehouse (CDW) that aggregates data from the university’s various clinical and administrative information systems, including the Epic (Epic Systems Corporation) electronic medical record (EMR) system. This study was approved by the UF Institutional Review Board (IRB201902362).

Study design
**Identify cancer patient cohorts**: Using UF IDR, we identified a general cancer (GC) cohort and a lung cancer screening (LCS) cohort. The GC cohort consists of 20,000 cancer patients sampled using a stratified random sampling (by cancer types), which has a total number of 1.5 million clinical notes. The LCS cohort was identified using the following rule-based phenotyping method:

1. The age at the first low-dose computed tomography (LDCT) date is between 50 and 80;
2. Most recent smoking status is Never smoker, but there’s evidence on patient ever smoked within the timeframe OR Most recent smoking status is Current smoker OR most recent smoking status is Former smoker, and at least one most recent quit year in structured data or note <= 15 years OR most recent smoking status is Former smoker but has Current smoker record in structured data within the timeframe; and

The LCS cohort consists of a total of 864 lung cancer patients with a total number of 161,933 various types of clinical notes. There are no patients existed in both cohorts.

**Training and test datasets**: We used the GC cohort to train various NLP methods, optimize model parameters. To develop a training corpus for SBDoH, we developed a filtering pipeline using note types (identified by domain experts) and a total number of 30 keywords to identify clinical notes potentially with rich content of SBDoH concepts. Using the filtering pipeline, we identified a total of 225,441 notes potentially with rich SBDoH concepts from the GC cohort, where 500 notes were randomly sampled for annotation. Two annotators manually identified all SBDoH concepts according to predefined guidelines. After annotation, we divided the 500 notes into a training set of 400 notes – used to develop various NLP models, and a test set of 100 notes – used to evaluate the performance of NLP models and select the best model. We trained various transformer-based NLP models using the training set and evaluated the performance using the test set (in terms of F1-score). Then, we identified the best NLP method according to the performance on test set and applied this model to extract SBDoH concepts from all clinical notes of the LCS cohort. To compare the NLP extracted SBDoH with those captured in structured EHRs, we developed a normalization pipeline to align the NLP results into predefined categories from the structured EHRs. Figure 1 shows an overview of the study design.

**Figure 1.** An overview of study design.

**NLP methods**

We explored two state-of-the-art transformer-based NLP methods, including (BERT)\(^7\) and (RoBERTa)\(^8\), and compared them with a widely used deep learning model based on Long short-term memory (LSTM)\(^9\). Previously, we have developed clinical transformer package\(^10\) based on the transformer architectures implemented in the HuggingFace\(^11\) in PyTorch\(^12\). We used the BERT model and RoBERTa model implemented in our clinical transformer package to develop transformer-based NLP solutions for SBDoH as they achieved the top 2 best performances in our previous study\(^13\). For BERT, we used the ‘base’ setting in this study. Following our previous studies\(^14,15,16\) on clinical transformers, we also examined pre-trained transformers from general English corpus (denoted as ‘_general’, e.g., ‘BERT_general’) and clinical transformers pre-trained using clinical notes from the Medical Information Mart for Intensive Care III (MIMIC-III) database\(^17\) (denoted as ‘_mimic’, e.g., ‘BERT_mimic’). We adopted the default parameters optimized in our clinical transformer package\(^18\).

**SBDoH from structured EHRs**
We collected 5 categories of structured SBDoH from the structured tables in the UF IDR database. At UF health IDR database, gender and ethnicity were captured as patient-level information, whereas the smoking status was captured for each encounter (one patient may have multiple smoking status from different encounters). For education and employment, there is no information captured in structured tables.

**NLP results alignment**

Our NLP system identified SBDoH concepts at the document-level, which is different from the structured SBDoH information that was captured at patient-level or encounter-level. In order to compare NLP results with structured EHRs, we developed a rule-based pipeline to align the document-level NLP results to structured EHRs. More specifically, we adopted a majority voting strategy to aggregate multiple mentions of gender and ethnicity from clinical notes into patient-level categories - the most frequent category will be used. Smoking status is a factor that may change over time, therefore, we compared smoking status at a “per patient / per year” setting. Under the “per patient / per year” setting, a patient is supposed to have a smoking status datapoint for each year in the EHR database – e.g., a patient with 5-year history is supposed to have 5 smoking datapoints, one for each year. We aggregated smoking information from both the structured EHRs (encounter-level) and NLP (document-level) into the “per patient/per year” setting for comparison. Similarly, the majority voting strategy was used. According to the structured smoking categories defined in structured EHRs, there are 7 different smoking categories. To align the NLP extracted smoking information to the 7 structured categories, we first generated a list of unique smoking status captured by NLP and then manually reviewed them (a total of 429) to map them into one of the 7 structured categories. As employment status and education were not captured in structured EHR, we were not able to compare.

**Evaluation**

We used both strict (i.e., the beginning and end boundaries of a concept have to be exactly the same with gold-standard annotation) and lenient precision, recall, and F1-score to evaluate our NLP systems. Precision is defined as (the number of SBDoH concepts correctly identified by the NLP system) / (total number of concepts identified by NLP); recall is defined as (the number of SBDoH concepts correctly identified by the NLP system) / (total number of concepts annotated by experts); F1-score is defined as “(2*precision*recall)/(precision+recall)”. To compare NLP results with structured data, we report the number of patients with structured SBDoH vs the number of patients with NLP extracted SBDoH; the number of patients who only have structured information vs the number of patients who only have NLP extracted information.

**Results**

Two annotators annotated 1,876 SBDoH concepts from 500 clinical notes. The inter-annotator agreement measured by token level kappa score with 40 overlapped clinical notes was 0.97, indicating the two annotators have a good agreement in the annotation. Table 1 shows the distribution of SBDoH concepts in the training and test set.

<table>
<thead>
<tr>
<th>Entity type</th>
<th>Training set</th>
<th>Test set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>441</td>
<td>120</td>
</tr>
<tr>
<td>Ethnicity</td>
<td>21</td>
<td>6</td>
</tr>
<tr>
<td>Smoking</td>
<td>665</td>
<td>178</td>
</tr>
<tr>
<td>Employment</td>
<td>27</td>
<td>10</td>
</tr>
<tr>
<td>Education</td>
<td>335</td>
<td>73</td>
</tr>
</tbody>
</table>

Table 2. Comparison of performance for BERT, RoBERTa, and LSTM-CRFs on the test set.

<table>
<thead>
<tr>
<th>Model</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT_general</td>
<td>0.8848</td>
<td>0.8734</td>
<td>0.8791</td>
<td>0.9058</td>
<td>0.8941</td>
<td>0.8999</td>
</tr>
</tbody>
</table>
Table 3. Detailed performance for each SBDoH category for the best NLP model - BERT_general.

<table>
<thead>
<tr>
<th>BERT_general</th>
<th>Strict</th>
<th>Lenient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td>Gender</td>
<td>0.9091</td>
<td>0.9167</td>
</tr>
<tr>
<td>Ethnicity</td>
<td>0.8571</td>
<td>1.0000</td>
</tr>
<tr>
<td>Education</td>
<td>0.8857</td>
<td>0.8493</td>
</tr>
<tr>
<td>Smoking</td>
<td>0.8764</td>
<td>0.8764</td>
</tr>
<tr>
<td>Employment</td>
<td>0.6667</td>
<td>0.4000</td>
</tr>
<tr>
<td>Overall</td>
<td>0.8848</td>
<td>0.8734</td>
</tr>
</tbody>
</table>

Table 2 compares two transformer-based NLP methods with a widely used deep learning model – LSTM-CRFs in extracting SBDoH concepts from clinical narratives. Both two transformer-based NLP methods outperformed LSTM-CRFs. Among the two transformer-based NLP methods, the BERT_general model achieved the best strict/lenient F1-score of 0.8791 and 0.8999 on the test set, respectively. Table 3 shows the detailed performance for each of the 5 SBDoH categories for the best NLP model. We applied the best NLP model, BERT_general, to extract SBDoH concepts from all clinical notes of the LCS patient cohort and compared them with structured EHRs. Table 4 shows the comparison results between NLP and structured EHRs. For the 864 patients in LCS cohort, the structured EHRs covered 99.65% patients for gender and ethnicity. However, the structured EHRs only covered 56.55% of smoking datapoints vs 71.57% from NLP. There is no information captured in structured EHRs for education and employment. Information about education and employment was documented in the narrative clinical text (39.35% for education and 47.22% for employment).

Table 4. Comparison of NLP extracted concepts with structured concepts for LCS cancer patients

<table>
<thead>
<tr>
<th>#concepts detected by NLP</th>
<th>Gender</th>
<th>Ethnicity</th>
<th>Smoking*</th>
<th>Education</th>
<th>Employment</th>
</tr>
</thead>
<tbody>
<tr>
<td>#patients with NLP detected concepts</td>
<td>88,015</td>
<td>14,866</td>
<td>104,201</td>
<td>22,460</td>
<td>2,236</td>
</tr>
<tr>
<td>#patients with structured concepts</td>
<td>861(99.65%)</td>
<td>713(82.52%)</td>
<td>5,736(71.57%)</td>
<td>340(39.35%)</td>
<td>408(47.22%)</td>
</tr>
<tr>
<td>#patients with NLP consistent with structured concepts</td>
<td>860(99.53%)</td>
<td>703(81.37%)</td>
<td>3,015(37.62%)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>#patients only have NLP concepts</td>
<td>2(0.23%)</td>
<td>0</td>
<td>1,517(18.92%)</td>
<td>340(39.35%)</td>
<td>408(47.22%)</td>
</tr>
</tbody>
</table>
When compared using from Table 4 that much of the detailed smoking datapoints were still documented in clinical text than structured EHRs with programs and lung cancer screening program compliance. The increase was likely of patients with smoking information in structured EHRs. From Figure 2 we can see that smoking information is consistently documented in clinical text, whereas, the proportion of patients with smoking information in one year (defined as “[number of patients with smoking information in one year] / [total number of patients in that year]”) and plotted the curve from 2009 to 2020 in Figure 2 (structured smoking information showed up in the LCS cohort starting from 2009; the EHR data for 2020 is not complete as we queried data in mid 2020). From Figure 2 we can see that smoking information is consistently documented in clinical text, whereas, the proportion of patients with smoking information in structured EHRs was low in the beginning but increased a lot over time. This increase was likely a result of institutional requirements for tobacco use documentation as part of quality assurance programs and lung cancer screening program compliance. Note that Figure 2 only shows the proportion of patients with smoking information or not, it can’t be used to assess how complete it is in structured EHRs. In fact, we can see from Table 4 that much of the detailed smoking datapoints were still documented in clinical text than structured EHRs when compared using the “per patient / per year” setting.

| #patients only have structured concepts | 2(0.23%) | 150(17.36%) | 308(3.84%) | 0 | 0 |

*For smoking, the numbers and proportions were calculated using datapoints – the 864 patients are supposed to have 104,201 datapoints using “per patient / per year” setting. The proportions for other SBDoH concepts were calculated according to a total number of 864 patients in LCS cohort.

**Discussion and conclusion**

SBDoH are important factors associated with cancer risk, prevention, screening, and survival. This study applied state-of-the-art transformer-based NLP methods to extract SBDoH concepts from clinical narratives and took lung cancer as a study case to compare the difference between clinical narratives and structured EHRs. Our experimental results show that two transformer-based NLP methods, BERT and RoBERTa, outperformed a widely used deep learning model – LSTM-CRFs – in extracting SBDoH from clinical narratives. Among the 4 transformer models, the BERT_general model pretrained using general English corpora achieved the best strict/lentent F1-score of 0.8791 and 0.8999, respectively, indicating the efficiency of transformer-based NLP methods. The BERT_general model pretrained using general English corpora outperformed another BERT_mimic model pretrained using MIMIC III clinical text, which is consistent with our previous study of using transformer models for de-identification of clinical notes.46 Similar to the protected health information in the de-identification task, the SBDoH concepts are closer to the general English language than the medical language. Therefore, the transformer models pretrained with a large volume of general English text are better in recognizing those SBDoH concepts closer to general English.

We examined 5 categories of SBDoH concepts in a lung cancer patient cohort. The best NLP model, BERT_general, was used to extract SBDoH concepts from all clinical notes of the LCS patient cohort. The comparison between NLP extracted results and structured EHRs shows that the structured EHRs have good coverage (>99% patients) for gender and ethnicity; but for smoking, education, and employment, much detailed information was documented in clinical narratives. For smoking, 71.57% of datapoints were documented in clinical notes, whereas, only 56.44% were captured in structured EHRs. Between the smoking datapoints extracted by NLP and structured EHRs, there are only 37.62% consistent with each other; there are 18.92% of smoking datapoints only from NLP, indicating the necessity of using both NLP and structured EHRs for cancer-related studies. For education and employment, currently there was no information captured in structured EHRs; but we can get 39.35% education information and 47.22% employment information from clinical notes using NLP. Smoking, education, and employment are important factors associated with many cancers. NLP systems could fill the gap of using these SBDoH factors in clinical studies. Smoking is an important BDoH factor for lung cancer. To better understand how smoking was documented in clinical narratives and structured EHRs over time, we calculated the normalized proportion of patients with smoking information for each year (defined as “[number of patients with smoking information in one year] / [total number of patients in that year]”) and plotted the curve from 2009 to 2020 in Figure 2 (structured smoking information showed up in the LCS cohort starting from 2009; the EHR data for 2020 is not complete as we queried data in mid 2020). From Figure 2 we can see that smoking information is consistently documented in clinical text, whereas, the proportion of patients with smoking information in structured EHRs was low in the beginning but increased a lot over time. This increase was likely a result of institutional requirements for tobacco use documentation as part of quality assurance programs and lung cancer screening program compliance. Note that Figure 2 only shows the proportion of patients with smoking information or not, it can’t be used to assess how complete it is in structured EHRs. In fact, we can see from Table 4 that much of the detailed smoking datapoints were still documented in clinical text than structured EHRs when compared using the “per patient / per year” setting.
Figure 2. A comparison of patients’ smoking status between clinic narratives and structured EHRs.

NLP is the key technology to extract SBDoH concepts from clinical narratives. Yet, most of the NLP studies focused on clinical factors such as diseases, medications; the application of NLP to extract SBDoH concepts is under-studied. Clinical NLP systems that can accurately identify SBDoH concepts are needed to fill the gap of using these factors in various health outcomes studies. The ultimate goal of this project is to develop a clinical NLP package that can accurately identify a wide range of SBDoH concepts from clinical narratives and populate them into structured EHR databases such as the NLP tables defined in the OHDSI common data model (CDM) and PCORnet CDM.

Limitations and future plans

This study has limitations. As a preliminary study, we limited SBDoH concepts to 5 categories including gender, ethnicity, smoking, education, and employment. Future studies need to examine more SBDoH categories associated with health outcomes. We took lung cancer as a study case to examine how current NLP models could identify SBDoH concepts and how complete they were captured in NLP vs structured EHRs. The documentation of SBDoH concepts might be different in other disease domains. During the time of this study analysis, the institutional policy changed to require more discrete data collection related to smoking use, which can impact the results. We expect more studies to examine the use of SBDoH in various diseases. Also, this study examined clinical notes from a single site, future work should examine the proposed transformer-based NLP system at cross-institute settings.

Acknowledgement

This study was partially supported by a Patient-Centered Outcomes Research Institute® (PCORI®) Award (ME-2018C3-14754), a grant from National Institute on Aging IR56AG 069880, a grant from the National Cancer Institute, 1R01CA246418 R01, a grant from CDC (Centers for Disease Control and Prevention) 1U18DP006512-01, the University of Florida (UF) SEED Program (DRPD-ROF2020, P0175580), and the Cancer Informatics and eHealth core jointly supported by the UF Health Cancer Center and the UF Clinical and Translational Science Institute. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding institutions.

References


Validation of Administrative Coding and Clinical Notes for Hospital-Acquired Acute Kidney Injury in Adults

Jianqiu Zhang, MPH1, Paul E. Drawz, MD2, Ying Zhu, PhD1, Gretchen Hultman, MPH1, Gyorgy Simon, PhD1, Genevieve B. Melton, MD, PhD1,3
1Institute for Health Informatics, 2Department of Medicine, 3Department of Surgery, University of Minnesota Minneapolis, MN, USA;

Abstract

Acute kidney injury (AKI) is potentially catastrophic and commonly seen among inpatients. In the United States, the quality of administrative coding data for capturing AKI accurately is questionable and needs to be updated. This retrospective study validated the quality of administrative coding for hospital-acquired AKI and explored the opportunities to improve the phenotyping performance by utilizing additional data sources from the electronic health record (EHR). A total of 34570 patients were included, and overall prevalence of AKI based on the KDIGO reference standard was 10.13%. We obtained significantly different quality measures (sensitivity:0.486, specificity:0.947, PPV:0.509, NPV:0.942 in the full cohort) of administrative coding from the previously reported ones in the U.S. Additional use of clinical notes by incorporating automatic NLP data extraction has been found to increase the AUC in phenotyping AKI, and AKI was better recognized in patients with heart failure, indicating disparities in the coding and management of AKI.

Introduction

Acute kidney injury (AKI) is a potentially catastrophic complication characterized by rapid loss of kidney function1. The prevalence of hospital-acquired AKI (HA-AKI) is high, affecting 21.6% of hospitalized adults and 33.7% of hospitalized children worldwide2. In the United States, although the incidence rates are lower, age-standardized rate of HA-AKI has increased by more than 139% over the past decade, accounting for 4.4% Medicare fee-for-service beneficiaries in 20183.

Given the incremental public health burden, AKI has received increasing attention in recent years. Healthcare data, including administrative data and clinical data, contains a wealth of information and can afford investigators the opportunity to study AKI on a large scale4. However, investigators need to be aware of the quality and limitations of these datasets 5,6. On one hand, although claims data has long been used by researchers and policy makers, the coding system was designed for administrative purposes and may not reflect the actual status of a patient7. Recent validation studies have been performed in some western countries and suggested that AKI diagnostic codes had very low sensitivity (< 40%) but relatively high specificity (> 90%)8. On the other hand, as the electronic health record (EHR) is emerging as a rich source of information for secondary use, it still requires substantial efforts to convert EHR data into a research database. To counter the challenges, a variety of EHR data components (i.e. clinical notes, lab values, medications) have been used for disease phenotyping, and many previous works and applications have shown promising results9. More specifically regarding the diagnosis of AKI, Grams et al and Ko et al examined the performance of manual chart review in phenotyping AKI. However, their results were not consistent, and their methods were human labor intensive and could be only conducted in small sample size10,11. Although there have been numerous efforts in studying automatic methods in extracting information associated with target diseases from unstructured EHR data, little been applied in the area of AKI.

Despite a large body of research in this field, there are significant gaps in the literature regarding validation of the current healthcare system in capturing HA-AKI. To the best of our knowledge, most validation studies in the U.S. were conducted in early 2000s but almost none have been reported since the publication of KDIGO diagnostic criteria, which is the first international clinical practice guideline on AKI12. Such quality measures (e.g. sensitivity, specificity) are underreported and needs to be updated. Meanwhile, little is known about the disparities in the coding performance across different spectrum of patients, or if leveraging EHR documentation and clinical data may help improve our ability to phenotype HA-AKI. In the study presented here, we utilized both claims and clinical data and evaluated the phenotyping performance of using one or more combinations of four major data components across pre-defined subgroups. We hypothesized that the additional information from EHR would be significant in improving the performance of AKI recognition, and quality performance of administrative coding in capturing AKI would be differed by subgroups.
Methods

Study setting and dataset

A retrospective study was conducted using data from the University of Minnesota Clinical and Translational Institute (CTSI). We extracted data from CTSI clinical data repository, which contains EHR data of > 2 million patients seen at 8 hospitals and more than 40 clinics in Fairview Health Services in Minnesota, USA beginning in 2005. This data repository incorporates administrative and clinical data and is divided into multiple categories regarding demographics, diagnoses, procedures, lab tests, vitals, medications and more. All structured data were de-identified and other free-text data such as clinical notes that could not be de-identified were stored in Academic Health Center Information Exchange (AHC-IE) servers and only accessible by AHC devices, to protect patient confidentiality. The study was approved by the University of Minnesota Institutional Review Board. The overall study design is shown in Figure 1.

Figure 1. Flowchart for validating AKI incidence in administrative and clinical datasets

Criteria for AKI

In this study, Kidney Disease: Improving Global Outcomes (KDIGO) staging system was used as the reference standard to identify AKI cases. According to the KDIGO guidelines, AKI is defined by changes in either serum creatinine (SCr) or urine output (UO) and staged in severity, as described in Table 1. Considering the availability of biomarkers in our database and limitations of the UO criterion, only the SCr diagnostic criterion was used in this study.

Table 1. KDIGO definition of AKI

<table>
<thead>
<tr>
<th>Stage</th>
<th>Serum creatinine</th>
<th>Urine output</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.5 to 1.9 times baseline &lt;br&gt;or ≥0.3 mg/dl (≥26.5 μmol/l) increase</td>
<td>&lt;0.5 ml/kg/hour for 6 to 12 hours</td>
</tr>
<tr>
<td>2</td>
<td>2.0 to 2.9 times baseline</td>
<td>&lt;0.5 ml/kg/hour for ≥12 hours</td>
</tr>
<tr>
<td>3</td>
<td>3.0 times baseline or increase in serum creatinine to ≥4.0 mg/dl (≥353.6 μmol/l) &lt;br&gt;or initiation of renal replacement therapy &lt;br&gt;or in patients &lt;18 years a decrease in eGFR to &lt;35 ml/minute per 1.73 m²</td>
<td>&lt;0.3 ml/kg/hour for ≥24 hours &lt;br&gt;or anuria for ≥12 hours</td>
</tr>
</tbody>
</table>

Baseline SCr is a key measure in the SCr criterion, and SCr levels measured between 365- and 7-day prior admission are usually considered as baseline. However, baseline SCr is frequently missing in EHR. Although
there are imputation methods such as back-calculations using MDRD, Cockcroft-Gault equations and etc., most of them are creatinine-based and are only useful when renal function is stable and may not be generalizable14. To avoid biased estimation and inaccurate diagnosis, actual SCr levels measured before admission were required. There is no conclusion regarding the best way to estimate baseline SCr when multiple values are available. We adopted the most exclusive rule in this study: using multiple baselines (median and the most recent value) and comparing them with subsequent SCr values simultaneously, if one reached the SCr threshold according to the KDIGO definition, AKI would be considered as present.

Study cohort
We extracted EHR data, containing both administrative and clinical data, of all adult hospitalized patients (age 18-100) admitted for any reasons between 2008 and 2019. Patients with the following characteristics were included in the study cohort: (1) first-time admission only; (2) having pre-admission SCr available; (3) having admission SCr value available, defined as the first SCr level measured within 36hr after admission; (4) having at least one additional SCr measured during inpatient stay, defined as any subsequent SCr measures to the admission SCr; (5) no AKI on admission; and (6) no prior AKI, no history of dialysis or renal transplant therapy (RRT). A look-back period of 1 year from the date of hospital admission was used to determine baseline renal function and comorbidities. We defined four subgroups of interest based on their comorbid conditions of diabetes, hypertension, heart failure and chronic kidney disease that were presented before the discharge time of each hospitalization. ICD 9 and 10 diagnostic codes were used to identify subgroups (Supplementary Appendix Table S3).

Decision rules for AKI
Decision rules for phenotyping HA-AKI were derived using one or more combinations of the following EHR components (codes listed in Supplementary Appendix A): (1) Laboratory SCr, by applying KDIGO diagnostic criteria and was considered as the reference standard; (2) ICD diagnostic codes, including ICD 9 and 10 codes for AKI; (3) AKI-related procedures, including CPT and ICD9/10 procedure codes for dialysis and RRT as group A (usually be considered as severe forms of AKI), as well as procedure codes or lab orders for renal ultrasound and hydration management as group B; and (4) Mentions of AKI in clinical notes, by using an natural language processing (NLP) enabled search engine called NLP-PIER. Due to the interchangeable use of medical terminology and potentially hierarchical relationships between terms (e.g., acute kidney injury is equivalent to acute kidney disease; hemodialysis is one type of dialysis), we used AKI-related medical terms in various expressions as key words and concept unique identifiers (CUI) for semantic search. NLP-PIER also allows fuzzy match on keywords, by enabling related misspellings (through frequency, spellgen score and edit distance) and semantically related terms (through frequency, cosine similarity) that are word vector-based15. These terms were then manually selected based on domain knowledge. Simple negations were excluded using regular expressions on CUI medical terms. Related administrative codes and CUI terms are listed in Supplementary Appendix Table S1-2.

Statistical analysis
We calculated the cohort and subgroup-specific sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV) and receiver operating characteristic area under curve (ROC AUC) of each decision rule as compared to changes in SCr using the KDIGO diagnostic criteria as the reference standard. P-values were calculated based on AUC scores between different decision rules within and across subgroups. All analysis were conducted using the R software (R Foundation for Statistical Computing, Vienna, Austria).

Results
Study population and characteristics
We identified a total of 34570 patients and unique hospitalization between 2008 and 2019 that met the inclusion criteria. Patient selection diagram is shown in Figure 1. The missing rate of baseline SCr was 64%, higher than the previously reported rate of 25–50%10. The numbers of patients without AKI on admission using either median or the most recent value of pre-admission SCr value were similar. Baseline characteristics of patients are outlined in Table 2. The overall prevalence of HA-AKI (based on KDIGO) in the full cohort was 10.13%. 73.78% of the cohort were in young adulthood (18-35 years), and these patients had a higher prevalence rate of HA-AKI (11.11%) than other age groups. Hypertension was the most common comorbid condition in our cohort, and the incidence rates of HA-AKI were particularly high (~22%) in patients with heart failure and chronic kidney disease (CKD) as compared with other subgroups.
Figure 2. Patients selection flow. EHRs of all adult inpatients were extracted during 2008-2019. 42,612 patients met the SCr inclusion criteria, and the size of final cohort was 34,570.

Table 2. Baseline characteristics of patients. No HA-AKI: patients did not develop HA-AKI during hospitalization. HA-AKI: patients developed HA-AKI during hospitalization.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Cohort</th>
<th>No HA-AKI</th>
<th>HA-AKI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demographics</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>age (n(%))</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18-35</td>
<td>23,245 (67.24)</td>
<td>20,662 (66.50)</td>
<td>2,583 (73.78)</td>
</tr>
<tr>
<td>36-55</td>
<td>3,392 (9.81)</td>
<td>3,126 (10.06)</td>
<td>266 (7.60)</td>
</tr>
<tr>
<td>&gt;55</td>
<td>7,933 (22.95)</td>
<td>7,281 (23.43)</td>
<td>652 (18.62)</td>
</tr>
<tr>
<td>total</td>
<td>34,570 (100)</td>
<td>31,069 (100)</td>
<td>3,501 (100)</td>
</tr>
</tbody>
</table>

| Women (n(%))    | 18,003 (52.08) | 16,480 (53.04) | 1,523 (43.50) |
| Race (n(%))     |              |              |                |
| White           | 31,089 (89.93) | 28,005 (90.14) | 3,084 (88.09)  |
| Asian           | 753 (2.18)     | 672 (2.16)    | 81 (2.31)      |
African American 22,55 (6.52) 1,983 (6.38) 272 (7.77) 
Other 473 (1.37) 409 (1.32) 64 (1.83) 

**Comorbid conditions (n(%))**

- Diabetes 9,143 (26.45) 7,817 (25.16) 1,326 (37.87) 
- Hypertension 14,090 (40.76) 12,362 (39.79) 1,728 (49.36) 
- Heart failure 4,506 (13.03) 3,502 (11.27) 1,004 (28.68) 
- Chronic kidney disease 6,465 (18.70) 5,022 (16.16) 1,443 (41.22) 

### Diagnostic performance by decision rules and subgroups

The decision rules used for performance measure are listed in Table 3. Since our study interest was in identifying additional EHR components for quality improvement in phenotyping while maintaining the use of administrative coding, the traditional combination of billing codes for AKI, dialysis and RRT were included in all decision rules. In general, low sensitivity and PPV (~ 0.441) but high specificity and NPV were observed across all decision rules I-IV. No single decision rule consistently received the best scores in sensitivity, specificity, PPV and NPV. The most comprehensive decision rule IV offered the best sensitivity in cohort average (0.525) but its PPV was the lowest. Similarly, the CKD group using decision rule IV had the best sensitivity across all groups but with the lowest specificity and NPV. Trade-offs between sensitivity vs. specificity and PPV vs. NPV were also observed when validating decision rule I: while the highest specificity and PPV were found in groups using the traditional methods, the corresponding sensitivity and NPV scores were among the lowest ones. ROC AUC scores were calculated and pairwise compared with the traditional rule (decision rule I) within the group and the scores of the same decision rule in other groups. The average AUC score in the full cohort was 0.715. Within the full cohort, the addition of clinical notes (decision rule III) had increased the AUC by 0.002 with p-value = 0.05 and was considered to be marginal. Meanwhile, the addition of procedure codes for ultrasound and hydration management decreased the AUC score by 0.03 (p-value = 0.03). No statistically significant changes were observed between different decision rules within each subgroup.

AUC of all decision rules were significantly high in patients with heart failure, with the highest AUC scores (0.738–0.743) and small p-values (0.004–0.028) as compared with other groups. Although the AUC values were also high in diabetic patients (average 0.727), they were not found statistically different from others. Patients with heart failure had the largest AUC. Similar distribution was found when using other decision rules.

### Table 3. Performance measure of decision rules. PPV: positive predictive value. NPV: negative predictive value. RRT: renal replacement therapy. CKD: chronic kidney disease.

<table>
<thead>
<tr>
<th>Decision rule</th>
<th>Group</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>PPV</th>
<th>NPV</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>I.</td>
<td>Diagnostic codes</td>
<td>Full Cohort</td>
<td>0.486</td>
<td>0.947</td>
<td>0.509</td>
<td>0.942</td>
</tr>
<tr>
<td></td>
<td>+ Procedure_A codes (dialysis, RRT)</td>
<td>CKD</td>
<td>0.633</td>
<td>0.804</td>
<td>0.482</td>
<td>0.884</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Diabetes</td>
<td>0.545</td>
<td>0.912</td>
<td>0.513</td>
<td>0.922</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Heart failure</td>
<td>0.600</td>
<td>0.887</td>
<td>0.603</td>
<td>0.885</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hypertension</td>
<td>0.494</td>
<td>0.937</td>
<td>0.521</td>
<td>0.930</td>
</tr>
<tr>
<td>II.</td>
<td>Diagnostic codes</td>
<td>Full Cohort</td>
<td>0.517</td>
<td>0.909</td>
<td>0.390</td>
<td>0.944</td>
</tr>
<tr>
<td></td>
<td>+ Procedure_A codes + Procedure_B codes (ultrasound, hydration management)</td>
<td>CKD</td>
<td>0.652</td>
<td>0.767</td>
<td>0.446</td>
<td>0.885</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Diabetes</td>
<td>0.572</td>
<td>0.876</td>
<td>0.438</td>
<td>0.923</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Heart failure</td>
<td>0.619</td>
<td>0.858</td>
<td>0.555</td>
<td>0.887</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hypertension</td>
<td>0.520</td>
<td>0.904</td>
<td>0.432</td>
<td>0.931</td>
</tr>
<tr>
<td>III.</td>
<td>Diagnostic codes</td>
<td>Full Cohort</td>
<td>0.495</td>
<td>0.941</td>
<td>0.486</td>
<td>0.943</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CKD</td>
<td>0.643</td>
<td>0.795</td>
<td>0.474</td>
<td>0.886</td>
</tr>
</tbody>
</table>
hypoalbuminemia (<4 g/dL) in off-threshold levels of biomarkers protected during cardiac surgery was measurable. Cardiovascular events coding or management guidelines were identified that we knew little was known about these codes even decreased the performance with ~12% increase in false positive, suggesting that specifically designed for this project, performs the best when it was trained and built for a specific task.

An explanation for observed phenomena might influence disease phenotyping has been demonstrated and validated by human chart review in many historical EHR research. This may be caused by different sizes of false positive (FP) and false negatives (FN) when using different cohorts. The FPs and FNs were well balanced in our sample, but data to examine these factors and differences in Grams et al were not available.

Besides the comparison with the literature, we also evaluated the additional use of non-traditional EHR components to improve our ability to HA-AKI phenotyping and potentially bring new knowledge about the disease. We observed that the combination of clinical notes and billing codes for AKI, dialysis and RRT improved the ROC AUC score in capturing AKI, and it achieved the best AUC across all subgroups. However, the p-value was close to the cut-off significance level (α=0.05) and may not hold as a strong evidence. Given that the importance of clinical notes in disease phenotyping has been demonstrated and validated by human chart review in many historical EHR research, an explanation for observed phenomena could be that our methods in automatic extracting AKI-related terms was imperfect, which has been a long-standing problem in the secondary use of EHR. An NLP pipeline usually performs the best when it was trained and built for a specific task. Since the NLP tool used in our study was not specifically designed for this project, the semantic search might not be perfect in capturing AKI, which could also explain the relatively low number of AKI cases defined by clinical notes. In addition to EHR notes, we also measured the performance of adding potentially AKI-related procedures (e.g., renal ultrasound and hydration management) but did not find these codes to be helpful in quality improvement. In some groups, the addition of these codes even decreased the performance with ~12% increase in false positive, suggesting that these procedures were not specific indicators of AKI and tend to overestimate the incidence of AKI.

Little was known about the disparities in the coding performance across different spectrum of patients. In this study, we identified that HA-AKI patients with heart failure were better recognized by all decision rules than other groups. The significantly high AUC score in patients with heart failure may be due to a higher degree of adherence to AKI coding or management guidelines in cardiac department. The reported prevalence of cardiac surgery-associated AKI is particularly high (up to 30%) with substantial increase in morbidity and mortality, so that the development of AKI is likely to be closely monitored by cardiologists. Although there is no uniform guideline in managing cardiovascular events-related AKI, there have been extensive research in renal protective strategies that are measurable. For instance, the European Society of Intensive Care Medicine published recommendations to kidney protection during cardiac surgery from aspects of preoperative, intraoperative and postoperative. Details in specific threshold levels of biomarkers, procedures and more measures (e.g., using exogenous albumin to correct hypoalbuminemia (<4 g/dL) in off-pump CABG surgery) were described in each aspect, and such strategies can

<table>
<thead>
<tr>
<th>+ Procedure_A codes + notes</th>
<th>Diabetes</th>
<th>0.555</th>
<th>0.906</th>
<th>0.500</th>
<th>0.923</th>
<th>0.730</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heart failure</td>
<td>0.606</td>
<td>0.880</td>
<td>0.591</td>
<td>0.886</td>
<td>0.743</td>
<td></td>
</tr>
<tr>
<td>Hypertension</td>
<td>0.504</td>
<td>0.931</td>
<td>0.503</td>
<td>0.931</td>
<td>0.717</td>
<td></td>
</tr>
<tr>
<td>IV. Diagnostic codes + notes</td>
<td>Full Cohort</td>
<td>0.525</td>
<td>0.903</td>
<td>0.379</td>
<td>0.944</td>
<td>0.714</td>
</tr>
<tr>
<td>Heart failure</td>
<td>0.625</td>
<td>0.852</td>
<td>0.547</td>
<td>0.888</td>
<td>0.738</td>
<td></td>
</tr>
<tr>
<td>Hypertension</td>
<td>0.528</td>
<td>0.899</td>
<td>0.422</td>
<td>0.932</td>
<td>0.714</td>
<td></td>
</tr>
</tbody>
</table>

**Discussion**

This study measured the quality performance of multiple decision rules using different EHR components in the task of phenotyping HA-AKI. As stated earlier, AKI is a common complication among hospitalized patients, and we found the prevalence of HA-AKI was 9.35% defined by common billing codes, near 8 times of the national reported one (1.17%) in 2018, but less than half of the rate (24.5%) from a meta-analysis in 2014. The difference may be due to the use of different data sources (Fairview EHR vs. The National Inpatient Sample (limited to billing codes) vs. systematic review) and different patient selection criteria. We also found that the incidence rates of HA-AKI were particularly high among patients with heart failure and CKD as expected. To the best of our knowledge, Grabs et al was the most recent validation work done in the U.S, and their sensitivity of billing code-defined AKI was 17.2%, which was substantially lower than our results (sensitivity: 48.6% - 66.1%). However, as Grabs et al observed, the sensitivity of billing codes was significantly higher in more recent time period. Since our study enrollment window was almost 11 years after Grabs et al, the changes in administrative coding and clinical practice over time might influence the performance of billing codes. Also, our PPV values were much lower than Grabs et al. This may be caused by different sizes of false positive (FP) and false negatives (FN) when using different cohorts. The FPs and FNs were well balanced in our sample, but data to examine these factors and differences in Grabs et al were not available.

Besides the comparison with the literature, we also evaluated the additional use of non-traditional EHR components to improve our ability to HA-AKI phenotyping and potentially bring new knowledge about the disease. We observed that the combination of clinical notes and billing codes for AKI, dialysis and RRT improved the ROC AUC score in capturing AKI, and it achieved the best AUC across all subgroups. However, the p-value was close to the cut-off significance level (α=0.05) and may not hold as a strong evidence. Given that the importance of clinical notes in disease phenotyping has been demonstrated and validated by human chart review in many historical EHR research, an explanation for observed phenomena could be that our methods in automatic extracting AKI-related terms was imperfect, which has been a long-standing problem in the secondary use of EHR. An NLP pipeline usually performs the best when it was trained and built for a specific task. Since the NLP tool used in our study was not specifically designed for this project, the semantic search might not be perfect in capturing AKI, which could also explain the relatively low number of AKI cases defined by clinical notes. In addition to EHR notes, we also measured the performance of adding potentially AKI-related procedures (e.g., renal ultrasound and hydration management) but did not find these codes to be helpful in quality improvement. In some groups, the addition of these codes even decreased the performance with ~12% increase in false positive, suggesting that these procedures were not specific indicators of AKI and tend to overestimate the incidence of AKI.

Little was known about the disparities in the coding performance across different spectrum of patients. In this study, we identified that HA-AKI patients with heart failure were better recognized by all decision rules than other groups. The significantly high AUC score in patients with heart failure may be due to a higher degree of adherence to AKI coding or management guidelines in cardiac department. The reported prevalence of cardiac surgery-associated AKI is particularly high (up to 30%) with substantial increase in morbidity and mortality, so that the development of AKI is likely to be closely monitored by cardiologists. Although there is no uniform guideline in managing cardiovascular events-related AKI, there have been extensive research in renal protective strategies that are measurable. For instance, the European Society of Intensive Care Medicine published recommendations to kidney protection during cardiac surgery from aspects of preoperative, intraoperative and postoperative. Details in specific threshold levels of biomarkers, procedures and more measures (e.g., using exogenous albumin to correct hypoalbuminemia (<4 g/dL) in off-pump CABG surgery) were described in each aspect, and such strategies can

<table>
<thead>
<tr>
<th>+ Procedure_A codes + notes</th>
<th>Diabetes</th>
<th>0.555</th>
<th>0.906</th>
<th>0.500</th>
<th>0.923</th>
<th>0.730</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heart failure</td>
<td>0.606</td>
<td>0.880</td>
<td>0.591</td>
<td>0.886</td>
<td>0.743</td>
<td></td>
</tr>
<tr>
<td>Hypertension</td>
<td>0.504</td>
<td>0.931</td>
<td>0.503</td>
<td>0.931</td>
<td>0.717</td>
<td></td>
</tr>
<tr>
<td>IV. Diagnostic codes + notes</td>
<td>Full Cohort</td>
<td>0.525</td>
<td>0.903</td>
<td>0.379</td>
<td>0.944</td>
<td>0.714</td>
</tr>
<tr>
<td>Heart failure</td>
<td>0.625</td>
<td>0.852</td>
<td>0.547</td>
<td>0.888</td>
<td>0.738</td>
<td></td>
</tr>
<tr>
<td>Hypertension</td>
<td>0.528</td>
<td>0.899</td>
<td>0.422</td>
<td>0.932</td>
<td>0.714</td>
<td></td>
</tr>
</tbody>
</table>
serve as strong references to follow in clinical practices\textsuperscript{20}. However, the complex mechanisms of other chronic conditions such as diabetes and hypertension may pose greater challenges in exploring a reliable and scalable way to manage AKI\textsuperscript{21}. For example, though there have been some guidelines published regarding the diabetes management in CKD patients (e.g., KDIGO 2020 Clinical Practice Guideline), few has been done in AKI\textsuperscript{22}. As mentioned earlier, a large portion (41.22\%) of HA-AKI patients in our study also had CKD, but the coding performance in CKD group did not show statistical differences from the diabetes or hypertension group, which was unexpected. CKD and AKI are known as highly inter-correlated diseases, so that CKD patients were supposed to be followed closely for changes in renal function and with better detection of AKI\textsuperscript{23,24}. This unexpected phenomenon may be caused by the limitations in the current criteria for nephrology referral. Oliva-Damaso et al questioned the weight of ‘age factor’ in the current nephrology referral criteria and found that younger CKD patients were less likely to be referred even their renal functions were poorer than the elderly\textsuperscript{25}. Since 73.78\% of HA-AKI patients in this study were in young adulthood, it is likely that some CKD patients did not received necessary nephrology consultation and had AKI under-recognized.

This study has several limitations. First of all, KDIGO diagnostic criteria is a silver standard and is imperfect\textsuperscript{13}. At present, glomerular filtration rate (GFR) is considered as the gold-standard diagnostic criteria for AKI, but GFR is almost never directly measured in the clinical setting. Instead, clinicians rely upon surrogate markers as markers of kidney dysfunction\textsuperscript{26}. Although KDIGO is widely adopted, using SCr as a biomarker of AKI has a number of drawbacks: (a) SCr level may be affected by factors other than renal function; (b) SCr measures the level of renal damage rather than renal function, and it may tend to underestimate AKI during early stages\textsuperscript{12,16}. However, we do believe that the advantages of using KDIGO outweigh the disadvantages. Second, patients without baseline SCr were excluded in this study, which consisted 63.69\% of the total number of inpatients in our study enrollment window. The exclusion on this population may introduce selection bias in study design. Third, subgroups were defined by billing codes and might not be accurate. It was a dilemma that although our study was focused on the capture of AKI, the attempt to explore the disparities by subgroups would inevitably introduce nested problems regarding the capture of conditions of interest (diabetes, hypertension, heart failure and CKD). Depending on the study questions, such inaccuracy may or may not be an issue, but it is important to acknowledge the level of possible errors/inaccuracy before drawing conclusions. This limitation also illustrated the importance of this work to facilitate other AKI research that will utilize admirative database and EHR components. Although the quality of billing codes capturing the subgroups in this database was unknown, our findings could still provide insights to identify certain groups of patients that need better or optimal management in AKI. Meanwhile, a 1-year lookback period for comorbidities might not be long enough, as relatively healthy patients might not have annual checkups and would not have related diagnostic codes recorded in their EHRs during that period. Fourth, our study cohort just represents the population enrolled in the Fairview Health Services system in Minnesota and almost 90\% of the cohort were White, and the majority of our study cohort were young adults, so that the results may not be generalizable. Last but not least, this study only utilized three EHR components (SCr lab value, billing code and clinical notes), but there may exist other EHR data elements that can leverage valuable information to AKI phenotyping.

In summary, we examined the performance of various EHR components in phenotyping AKI in this study. Our results not only updated our knowledge of the the quality of administrative coding for HA-AKI that is important to know for each end user, but also found that leveraging EHR documentation may help improve our ability to phenotyping HA-AKI. To the extent of our knowledge, our study was the first validation work of AKI that adopted automated data extraction from clinical notes. The results indicated clinical notes as a potential data source to improve the current administrative coding system, reduce incorrect coding and billing, and promote better communications between patients, insurers and providers. The improved performance also suggested the feasibility of this method in phenotyping AKI, which can facilitate researchers to study AKI with better quality and less additional cost as compared with human chart review. What’s more, the additional contents (i.e., results and findings from pathology report) and EHR elements (not limited to notes) could be used to discover a more comprehensive risk panel of AKI. Such risk panels may or may not serve as diagnostic evidence but can help discover additional measures to refine the current KDIGO criteria. Differences found across subgroups may be due to varying degrees of adherence to AKI coding or management guidelines. The results in discovering disparities by subgroups may help researchers to identify a specific population and build evidence-based predictive model that are optimized for that population, to ultimately improve patients’ health outcomes. Our results should be interpreted while taking into account the limitations inherent to our study design. In future studies, the diagnostic criteria for AKI needs to be further clarified and refined by using more unbiased imputation method when baseline is missing. Meanwhile, NLP
pipelines need to be optimized for AKI phenotyping. In addition to clinical notes, a panel of reliable indicators of AKI could also be used for better recognition of AKI.

**Conclusion**

Using common definitions of AKI and additional EHR data elements, we obtained significantly different quality measures from the previously reported ones in the U.S. This study highlighted the major limitations of administrative coding in capturing HA-AKI, as well as the opportunities for data quality improvement by potentially leveraging EHR data sources for end users, and ultimately predicting high risk of AKI to intervene earlier and decrease patient morbidity.

**References**


Acknowledgements
Funding source and assistant: Clinical Translational Science Award (NIH UL1TR000114), Agency for Healthcare Research and Quality (R01HS24532), the University of Minnesota Department of Surgery, Institute for Health Informatics, Academic Health Center-Information Exchange (AHC-IE), University of Minnesota Clinical and Translational Institute (CTSI), Fairview Health Services.

Supplementary Appendix

Table S1: Diagnostic codes and CUI terms for AKI

<table>
<thead>
<tr>
<th>Diagnosis</th>
<th>ICD-9 Codes</th>
<th>ICD-10 Codes</th>
<th>CUI</th>
</tr>
</thead>
<tbody>
<tr>
<td>AKI</td>
<td>584.x</td>
<td>N17.x</td>
<td>C0022660, C2349570, C2609414, C3854173, C4302681, C4302831, C4302832, C4512067, C4524217, C4524220, C4524222, C4552597, C4686607, C4687027, C5397148</td>
</tr>
</tbody>
</table>

Table S2: Administrative codes and CUI terms for AKI-related procedures

<table>
<thead>
<tr>
<th>Procedures</th>
<th>CPT Codes</th>
<th>ICD-9 Codes</th>
<th>ICD-10 Codes</th>
<th>CUI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dialysis</td>
<td>90935-91003</td>
<td>V45.1, V56, 39.95, 54.98</td>
<td>Z49, Z45.2, Z99.2</td>
<td>C0011945, C0011946, C0019004, C0031139, C0403464, C0455151, C0478581, C0842339, C0849861, C1161201, C1546731, C1997821, C2136533, C3516825, C3875178, C3877552, C4049377, C4316277, C455129</td>
</tr>
<tr>
<td>RRT</td>
<td>868, 50340, 50360, 50365</td>
<td>V42.0, 55.6x, 996.81</td>
<td>Z94.0, 0TY00Z0-1, 0TY10Z0-1, T86.1x</td>
<td>C0022671, C0041176, C0748332, C0948380, C0948351, C2063399, C2066095, C2170316, C2170328</td>
</tr>
<tr>
<td>Renal ultrasound</td>
<td>76770, 76775, 76776, 93975</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Hydration management</td>
<td>96360, 96361</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Diagnosis</td>
<td>ICD-9 Codes</td>
<td>ICD-10 Codes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------------------------</td>
<td>-------------</td>
<td>--------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diabetes</td>
<td>250.x</td>
<td>E10.x, E11.x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Heart failure</td>
<td>398.91, 402.x, 404.x, 428.x</td>
<td>I09.81, I11.x, I13.x, I50.x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hypertension</td>
<td>401.x, 401.9, 402.x, 403.x, 404.x, 405.x, 416.0, 416.8</td>
<td>I10, I11.x, I12.x, I13.x, I15, I15.x, I16.x, I27.0, I27.2, I27.20-I27.24, I27.29</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Abstract

Epilepsy is a common serious neurological disorder that affects more than 65 million persons worldwide and it is characterized by repeated seizures that lead to higher mortality and disabilities with corresponding negative impact on the quality of life of patients. Network science methods that represent brain regions as nodes and the interactions between brain regions as edges have been extensively used in characterizing network changes in neurological disorders. However, the limited ability of graph network models to represent high dimensional brain interactions are being increasingly realized in the computational neuroscience community. In particular, recent advances in algebraic topology research have led to the development of a large number of applications in brain network studies using topological structures. In this paper, we build on a fundamental construct of cliques, which are all-to-all connected nodes with a k-clique in a graph \( G(V, E) \), where \( V \) is set of nodes and \( E \) is set of edges, consisting of \( k \)-nodes to characterize the brain network dynamics in epilepsy patients using topological structures. Cliques represent brain regions that are coupled for similar functions or engage in information exchange; therefore, cliques are suitable structures to characterize the dynamics of brain dynamics in neurological disorders. We propose to detect and use clique structures during well-defined clinical events, such as epileptic seizures, to combine non-linear correlation measures in a matrix with identification of geometric structures underlying brain connectivity networks to identify discriminating features that can be used for clinical decision making in epilepsy neurological disorder.

Introduction

The characteristics of brain network connectivity representing both white matter fiber tracts (brain structural network) and interactions patterns between different brain regions (brain functional network) are being increasingly recognized as important biomarkers in a variety of neurological disorders [1-3]. In particular, network science methods based on graph theory, which represent brain regions as nodes and the connections between the brain regions as edges, have been used in a large number of studies to identify differences between brain networks in healthy subjects and patients with neurological disorders [3-5]. These studies have found that common characteristics of healthy brain networks, such as “small world network” structures, which include several local connections and few long-range connections across brain regions, are disrupted in patients with neurological disorders [3]. For example, network analysis of electroencephalography (EEG) recordings in epilepsy patients during seizures (ictal period) and between seizures (interictal period) have shown changes in brain connectivity using network metrics such as clustering coefficient, which measure local connectivity, and characteristic path length, which measure global connectivity [6].

Epilepsy is one of the most common serious neurological disorder that affects more than 65 million persons worldwide with more than 2.8 million persons diagnosed with epilepsy in the United States alone [7]. The application of network science method to EEG recordings has lead to greater understanding of epilepsy as a network-based neurological disorder and this has led to development of treatment methods that address seizure events as network events [8]. Further, the application of graph model-based network analysis techniques have also provided new data about comorbidities in epilepsy such as cognitive decline [9, 10]. However, the application of graph theory-based network analysis techniques in epilepsy has led to disparate results with some studies reporting an increase in connectivity [11, 12], while other studies have reported decreased connectivity [13, 14]. Although, these differences may be due to the study methodology, study population among other factors; however, there is growing recognition in the neuroinformatics community that graph models have inherent limitations due their binary representations.

Graph models can only represent interactions between two brain regions (binary interactions); therefore, they are severely constrained to represent high dimensional interactions that involve two or more brain regions simultaneously, which occur frequently as highlighted in epileptic seizure involving multiple foci of interactions [15, 16]. Further, graph-based models often use arbitrary threshold values to include or exclude interactions between brain regions as an edge between two nodes [15]. Therefore, recent studies in brain network research have characterized existing graph theory models as being oversimplification of brain interactions dynamics with a resulting lack of accuracy in representing interaction dynamics [16-18].
To address this critical limitation of graph-based models, we describe the development of a new Epilepsy Topology Data Analysis (ETDA) framework based on recent advances in the field of algebraic topology [19] to accurately characterize aberrant brain interactions in epilepsy patients. Figure 1 shows an overview of our approach using EEG data recorded from intracranial depth electrodes implanted in an epilepsy patient (as part of presurgical evaluation process). Using a nonlinear correlation coefficient measure $h^2$ to compute the degree of coupling between signals recorded by electrode contacts from two brain regions, we generate a directed graph network that represents binary relations between brain regions during epileptic seizure [20, 21]. The nonlinear correlation coefficient is based on the estimation of amplitude $y$ of signal $Y$ from the amplitude $x$ of signal $X$, such that the predicted value of $y$ for given value of $x$ is a regression curve [22]. It is important to note that the $h^2$ values are asymmetric, that is, it is possible that $h^2(y|x) \neq h^2(x|y)$, which can be used to infer directionality of the association between signal [22]. Figure 1 shows that a variety of graph motifs that are formed during a phase called ictal 1 in epileptic seizure when the epileptic seizure propagates beyond the initial seizure onset zone to additional brain regions (subsequent spread of epileptic seizure are labeled as ictal 2 phase, ictal 3 phase etc. before termination of seizure). Among the different graph motifs formed during the ictal period, we select cliques consisting of the set of all-to-all connected vertices in the directed graph model to compute the topological structure called a clique complex as shown in Figure 1 with electrode contacts labeled (RF1, RF2, and RF3) as well as (LF2, LF3, and LJ3) forming two cliques.

**Clique Topology of Epileptic Seizure Networks.** A clique motif in a graph network represents brain regions that have high degree of interaction, and they are of particular relevance in studying the onset and propagation of epileptic seizures. Epilepsy is characterized by a hypersynchronous state of the brain that results in formation of a regular network topology in epilepsy patients as compared to healthy subjects [3, 6, 8]. Therefore, clique motifs are representative of brain regions that have a high degree of interactions and are likely to play a key role in epileptic seizure networks. In this paper, we use the concept of simplex from algebraic topology, such that if the nodes of a graph are assumed to be in higher-dimensional space, then a clique of the graph represents a simplex given by the convex hull of the corresponding nodes [19, 23]. A simplex in algebraic topology is a generalization of graph model to incorporate multi-dimensional interactions (polyadic or n-ary relations) and the collection of these simplices form a simplicial complex that is analyzed to track the lifecycle of these topological structures during normal and abnormal brain activity [15, 18, 20].

We propose that these topological structures can successfully model high-dimensional interactions, such as epileptic seizure activities, that involve two or more brain regions simultaneously; therefore, they address the critical limitations of graph theory-based network analysis methods. Further, using well-defined algebraic topology methods such as persistent homology enables us to characterize local and global interactions between brain regions based on the occurrence of topological constructs called cycles and boundaries respectively [16]. The objectives of the ETDA framework are: (1) accurate modeling of high-dimensional interaction between brain regions during seizure activities using the set of clique motifs formed in epileptic seizure network models; (2) identify topological structure that have a long lifetime, which can be used to potentially differentiate brain regions that have high degree of participation in seizure activities over others; and (3) translate the summary statistics derived from ETDA to clinical research and patient care, such as improving the lateralization of the epileptogenic zone in patients refractory to medication during presurgical evaluation [24, 25]. The ETDA framework enables the development of a new approach for characterizing the complex dynamics of brain networks during epileptic seizure.

**Figure 1:** An overview of the Epilepsy Topological Data Analysis (ETDA) using clique complexes.
and also provides novel insights into the brain networks that underpin serious neurological disorders as compared to healthy individuals.

1.1 Background.

Directed network model of epileptic seizures from EEG data. Multiple signal analysis techniques have been developed to compute the coupling between signal data recorded from different brain regions, for example using linear cross-correlation and coherence functions. In particular, a method developed by Pijn et al. based on nonlinear regression coefficient \( h^2 \) has been widely used to effectively characterize the association between two EEG signals during seizure activity [22, 26, 27]. The directionality computed from EEG signal data is an important feature of the \( h^2 \) measure as it enables source localization in the epileptic network [26, 28]. The non-linear measure is maximized by taking into account all possible time lag values \( \tau \) between the two signals in the direction X to Y and using the maximum value of the measure, that is, \( h^2_{XY} = \max_{\tau_{\text{min}}<\tau<\tau_{\text{max}}} h^2_{XY} \). We use the resulting directed graph with nodes representing intracranial electrodes contacts and edges representing correlation (Figure 1) to compute simplicial complexes.

Topology data analysis. A clique in a graph epileptic seizure network graph \( SNG \) consists of a collection of all-to-all connected vertices and the set of all cliques in a graph \( SNG \) is called the clique complex of \( SNG \) and is written \( X(SNG) \) [18]. It is important to note that all subsets of a clique \( \Omega \) are also cliques, therefore \( X(SNG) = \{ \Omega \subset [M] | \Omega \text{ is a clique of } SNG \} \). Therefore, the cliques in \( X(SNG) \) consisting of \( n+1 \) nodes represent \( n \)-dimensional simplices [18]. The homology of the clique complex \( X(SNG) \) computes the relations between cliques in \( SNG \) and the homology counts the cycles in a clique complex. A cycle represents a collection of nodes which interact with one another and is a boundary when all of the nodes in that collection interact with all the other nodes. In this way, the homology of the clique complex represents higher-dimensional structures that are generalization of the number of cliques in a graph. The clique topology of a graph is computed using Betti numbers that track the dimensions of homology groups (we refer to the [23] and [29] for a detailed discussion on computation of Betti numbers). Therefore, the \( n \)-th Betti number of \( X(SNG) \), denoted by \( B_n(X(SNG)) \) is the rank of a \( k \)-vector space \( H_n(X(SNG);k) \) [18].

Summary Statistics of Topological Structures using Betti Numbers. The arrangement of cliques in an order complex of a matrix can be used to detect random or geometric structures. Clique topology studies the interactions (such as overlapping) between the cliques in the entire order complex. The topological structure of cliques in graph is quantified by “filling in” all cliques and counting non-contractible cycles (arrangement of cliques that bind the “holes”). A 1-cycle bounds 2D area, a 2-cycle bound a 3D volume. As the edge density \( p \) is varied, cycles may be created, destroyed, or different cycles might merge together. This information is tracked by computing the set of Betti numbers that count the independent \( n \)-cycles in each graph after all cliques have been filled in for a cycle. The Betti numbers across all graphs in an order complex yields a plot of the Betti curve.

2. Methods

There is an emerging consensus that the common characteristics of healthy brain networks such as a small world network structures are disrupted in neurological disorders; however, it is not clear how these disruptions occur and what is the effect of these disruptions on the core neurocognitive networks, such as the default mode network (DMN), salience network (SN), and central executive network (CEN), which are associated with memory, speech, or consciousness state [30-32]. The ETDA framework is an innovative approach to effectively address the challenges of: (1) Modeling the physiology of interactions in the aberrant brain networks to provide insights into how disruptions occur; and (2) Computing topological structures that are characteristic of clinical events to understand the effect of disruptions on neurocognitive functions. In addition to accurately modeling seizure networks, the clique topology model has applications in a variety of brain connectivity research, including: (1) identification of discriminating components of a seizure network for individual patients across multiple seizures to construct a patient-specific map of seizures; and (2) the metrics derived from the seizure network clique topology will be interpretable by clinicians for use at the point of care. In this section, we describe the approach used to compute the maximal simplices for the clique complexes formed during seizure events, the computation of Betti numbers, and bottleneck distance.

Study Data. We used a unique dataset of high-resolution stereotactic encephalograph (SEEG) recordings from a Level 4 epilepsy center at the University Hospitals Cleveland Medical Center (UH-CMC). The UH-CMC epilepsy center carries out pre-surgical evaluation of patients with drug-resistant epilepsy using multi-modal physiological parameters including EEG, EKG, video, autonomic, evoked potential, O2, CO2 and respiratory rate measurements. The EEG acquisition system is a Nihon Kohden (Japan) system, which is capable of simultaneously recording up to 192 channels of EEG at 1000 samples/second along with video recordings. The intracranial electrodes consist of platinum-iridium contacts with a diameter of 1.1 mm by 2.5 mm and each lead was 31 cm long with 10 - 12 electrodes that were placed every 5 mm, beginning from the tip. In addition to video-EEG recording, 16 analog input signals can be connected to the system. Table 1 gives the details of the five patients used in this study. Although multiple seizures were recorded for each patient, we selected two seizures with highest phase of ictal phases, that is, distinct phases in which seizure propagation was clinical identified to include additional brain regions [20].
<table>
<thead>
<tr>
<th>Patient ID</th>
<th>Demography</th>
<th>Etiology</th>
<th>Medications</th>
<th>Seizure Details</th>
<th>Seizure ID</th>
<th>Average Event Duration</th>
<th>Electrode Channel List</th>
</tr>
</thead>
<tbody>
<tr>
<td>NTHO</td>
<td>Male 29 y/o</td>
<td>Traumatic Brain Injury</td>
<td>Keppra; Depakote</td>
<td>Epileptogenic Zone: Right Temporal</td>
<td>Seizure 1 80 seconds</td>
<td>TT11, TT12, HH9, HH10, PT11, PT12, HB4, HB5, HB6, HB7, HB8, HB9, HH1, HH2, HH3, HH4, AM5, AM6, AM7, AM8, AM9, AM10</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Seizure 2 139.5 seconds</td>
<td>TT11, TT12, HH9, HH10, HB4, HB5, HB6, HB7, HB8, HB9, PT11, PT12, AM5, AM6, AM7, AM8, AM9, AM10, CG6, CG7, CG8, CG9, CG10</td>
<td></td>
</tr>
<tr>
<td>CASI</td>
<td>Male 24 y/o</td>
<td>Left Mesial Temporal Sclerosis</td>
<td>Levetiracetam; Lamotrigine; Lacosamide</td>
<td>Epileptogenic Zone: Left Mesial Temporal</td>
<td>Seizure 1 20.3 seconds</td>
<td>HB1, HB2, HB3, HH1, HH2, PC3, PC4, PC5, PC6, AG1, AG2, TP1, TP2, A11, A12, A13, A14, P11, P12, P13, P14, PT7, PT8, PT9, PT10, PT11, PT12</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Seizure 2 21.3 seconds</td>
<td>HB1, HB2, HB3, HH1, HH2, PC3, PC4, PC5, PC6, AG1, AG2, TP1, TP2, TP3, TP4, A11, A12, A13, A14, OF1, OF2, OF3, OF4, OF5, OF6, P11, P12, P13, P14, PT9, PT10, PT11, PT12</td>
<td></td>
</tr>
<tr>
<td>DIMX</td>
<td>Male 27 y/o</td>
<td>Unknown</td>
<td>Keppra; Lamictal</td>
<td>Epileptogenic Zone: Left Hippocampal</td>
<td>Seizure 3 90 seconds</td>
<td>HB1, HB2, HB3, HH1, HH2, TP1, TP2, AM1, AM2, AM3, AM4</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Seizure 4 79 seconds</td>
<td>HB1, HB2, HB3, HH1, HH2, TP1, TP2, TP3, TP4, AM1, AM2, AM3, AM4, OF3, OF4, OF5, OF6</td>
<td></td>
</tr>
<tr>
<td>HHUG</td>
<td>Male 39 y/o</td>
<td>Unknown</td>
<td>Zonisamide; Depakote; Clonazepam</td>
<td>Epileptogenic Zone: Left Mesial Temporal</td>
<td>Seizure 1 19.5 seconds</td>
<td>HH1, HH2, HH3, HH4, HB2, HB3, AM1, AM2, AM3, AM4, OF1, OF2, OF3, OF4, OF5, OF6, PT1, PT2, PT3, PT4</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Seizure 2 26.3 seconds</td>
<td>HH1, HH2, HH3, HH4, HB2, HB3, AM1, AM2, AM3, AM4, OF1, OF2, OF3, OF4, PT1, PT2, PT3, PT4, PC5, PC6, PC7, PC8, TT1, TT2, TT3, TT4, TT5, TT6, TT7, TT8, TT9, TT10</td>
<td></td>
</tr>
<tr>
<td>NCEN</td>
<td>Male 26 y/o</td>
<td>Unknown</td>
<td>Vimpat; Keppra; Depakote</td>
<td>Epileptogenic Zone: Left Mesial Temporal</td>
<td>Seizure 1 18.25 seconds</td>
<td>AM1, AM2, HH1, HH2, HB1, HB2, SC1, SC2, PT1, PT2</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Seizure 2 18.5 seconds</td>
<td>AM1, AM2, HH1, HH2, HB1, HB2, SC1, SC2, PT1, PT2</td>
<td></td>
</tr>
</tbody>
</table>

2.1 Computation of Clique Complexes from SEEG Data. The fundamental unit of study in algebraic topology is the simplex; multiple simplices are then brought together in a “controlled” way to form a simplicial complex – this “control” is achieved by positioning simplices not arbitrarily, but by gluing them together in such a way that any two simplices are either entirely disjoint or have an intersection that is itself a simplex. These ideas can be developed entirely in the abstract so that they can be adapted to a variety of different constructs (see [23] for full details), but we will focus in this paper on a specific notion of a simplicial complex that is very well suited for graphs, namely the clique complex [15]. As suggested by the name, we define a simplex in the clique complex of a graph to be a network clique, i.e., a complete subgraph, or a subset of the vertices and edges of the graph such that any two vertices in the subset are connected by some edge. More specifically, since simplices are associated with a non-negative integer called its dimension, we define a simplex of dimension n, or an n-simplex, to be a clique on n+1 vertices. Note how this corresponds to the “gluing rule” above: any subgraph of a clique is itself a clique, and therefore the intersection of any two cliques must be empty or a clique, and hence a simplex.
Given a graph, we can identify its simplices inductively, starting with the list of the graph’s 0-simplices (i.e., vertices). From a list of the graph’s n-simplices, we can construct the list of (n+1)-simplices by iterating over the list of n-simplices and the list of vertices and creating a corresponding (n+1)-simplex if a given vertex is connected to each node of a given n-simplex; given the graph as an adjacency matrix, this lookup can be done instantly. This procedure explicitly illustrates how lower-dimensional simplices can be subsumed by higher-dimensional ones, and we would like to focus on the higher-dimensional simplices that demonstrate a greater interconnectivity of nodes. Therefore, at each step of the process above, we will also be discarding those n-simplices that are contained in a formed (n+1)-simplex (these n-simplices are traditionally termed the faces of the (n+1)-simplex). This gives a final list of lists of maximal simplices for the clique complex.

2.2 Computation of Betti Numbers. The homology of a clique complex computes the relations between cliques and the homology counts the cycles in a clique complex. This represents higher-dimensional structures that are generalization of the cliques in a graph. The clique topology of a graph is computed using Betti numbers that track the dimensions of homology groups (we refer to the [23] and [29] for a detailed discussion on computation of Betti numbers). Relative ordering of entries in a correlation matrix, $X_{ij} < X_{kl}$ whenever $Y_{ij} < Y_{kl}$ is called an order complex $\text{ord}(X)$ [18]. Order complex represented as a nested sequence of graphs where each subsequent graph includes additional edges corresponding to the next largest matrix entry $X_{ij}$. Any quantity computed from order complex is invariant under transformation as $\text{ord}(X) = \text{ord}(Y)$. The arrangement of cliques in an order complex of a matrix can be used to detect random or geometric structures. Clique topology is a measure of how cliques fit together and overlap across entire order complex. Topological structure of cliques in graph is quantified by “filling in” all cliques and counting non-contractible cycles (arrangement of cliques that bind the “holes”). A 1-cycle bounds 2-D area, a 2-cycle bound a 3-D volume. As the edge density $p$ increases, new cycles are created, modified, and eventually destroyed, which are tracked by computing set of Betti numbers that count the independent n-cycles in each graph after all cliques have been filled in for a cycle. The Betti numbers across all graphs in an order complex yields a plot of Betti curves. In the Results section, we illustrate the Betti curves computed from the five patients used in this study.

2.3 Persistence Diagram to Characterize Topological Structure Across Seizure Events. The lifetime of a topological structure can be plotted over filtration in horizontal and vertical axes to generate a persistence diagram [33]. A point at $(x,y)$ represents a generator born at an edge density of $x$ and filled in at an edge density of $y$. Since a generator cannot die before it is born, $y$ is always greater than or equal to $x$, so we draw the $y=x$ diagonal line (where $y-x=0$ persistence) and use the distance of a point from the diagonal as a handy heuristic to estimate persistence. If a point is close to or on the diagonal, it did not persist long after first appearing and can almost be treated as noise. If a point is far from the diagonal, the point persisted long after being born (i.e., many edges had to be added before it could be filled in). In the latter case, keep in mind that edges are added in order of weight, so persistence does not say anything about the size of a hole: for example, a small hole such as a square that is only two edges away from forming a clique/pyramid may prove to be "significant" in that there really aren't any weighty edges that could fill it in. In summary, persistence diagrams make important generators stand out from the rest at a glance.

After the generation of the persistence diagrams, we use the bottleneck distance measure, which characterizes the difference between two persistence diagrams. Therefore, bottleneck distance is used to find the best bijective matching between points of two persistence diagrams, that is, the best way to pair points while minimizing distance between them. The bottleneck distance between A and B can be thought of as assuming diagram B is an approximation (or noisy version) of diagram A and measuring the error in the approximation. In our table, there are blanks whenever the current phase has trivial homology (i.e., no points in the persistence diagram). If two diagrams have low bottleneck distance, it means the distribution of their generators/points is very similar. This is extremely useful for testing hypotheses: if we generate a diagram A which is known to uniquely describe a graph with a certain property P and we find that diagram B is at a low bottleneck distance from A, we can say with some confidence that the graph upon which diagram B was generated also has the property P. Likewise, a high bottleneck distance would support the claim that the graph of B does not have property P. This technique can also be used to show that our diagrams do not have the characteristics of diagrams generated from random, unstructured data. We discuss the results of computing the Betti numbers and generation of persistent homology diagrams from seizure events from the epilepsy patients in the following section.

3. Results

The objective of our evaluation approach is to characterize the lifecycle of clique complexes computed from EEG data recorded from the five epilepsy patients across different phases of epileptic seizure. In the first section, we describe the persistence diagrams for the patients and in the second section we discuss the bottleneck distance computed from the persistence diagrams.
3.1 Persistence diagram derived from EEG data in five patients. Due to space constraints, we illustrate the persistence diagrams for two patients corresponding to two seizure events with multiple phases, including seizure onset and ictal phases.

Figure 2: Persistence diagram for patient ID HHUG across multiple seizure phases during two distinct seizure events.
We also illustrate the persistence diagram of the second patient (study ID: DIMX) across different dimensions. It is interesting to note that no topological features were detected in the second seizure event for dimension = 3.

Figure 3. Persistence diagram for patient ID DIMX across multiple seizure phases in two distinct seizure events.
3.2 Bottleneck distance computed over the persistence diagrams in five patients. As we discussed in the Method section, bottleneck distance is used to compute the difference between two persistence diagrams corresponding to distinct seizure phases. In Table 2 we describe the bottleneck distance computed for different seizure stages across different seizures in each of the five patients.

Table 2: Bottleneck distance computed for the five patients across different phases of epileptic seizure.

<table>
<thead>
<tr>
<th>Patient Study ID</th>
<th>Seizure ID</th>
<th>Dimension Value</th>
<th>Seizure Onset</th>
<th>Ictal 1 Phase</th>
<th>Ictal 2 Phase</th>
<th>Ictal 3 Phase</th>
<th>Ictal 4 Phase</th>
</tr>
</thead>
<tbody>
<tr>
<td>NCEN</td>
<td>1</td>
<td>0</td>
<td>0.31081081081081100</td>
<td>0.10810810810810800</td>
<td>0.2567567567567570</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NCEN</td>
<td>1</td>
<td>1</td>
<td>0.198796729678615200</td>
<td>0.31291946308724800</td>
<td>0.31291946308724800</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NCEN</td>
<td>2</td>
<td>0</td>
<td>0.431255167758253490</td>
<td>0.12322278815166000</td>
<td>0.02843601895734600</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NCEN</td>
<td>2</td>
<td>1</td>
<td>0.508672037914692900</td>
<td>0.13744075829383900</td>
<td>0.08530856872303790</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NCEN</td>
<td>2</td>
<td>2</td>
<td>0.094240837696353510</td>
<td>0.08376963507853400</td>
<td>0.099476439797057590</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NTHO</td>
<td>1</td>
<td>0</td>
<td>0.036450268456375800</td>
<td>0.31291946308724800</td>
<td>0.31291946308724800</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NTHO</td>
<td>1</td>
<td>1</td>
<td>0.056872037914692900</td>
<td>0.13744075829383900</td>
<td>0.08530856872303790</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NTHO</td>
<td>1</td>
<td>2</td>
<td>0.094240837696353510</td>
<td>0.08376963507853400</td>
<td>0.099476439797057590</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NTHO</td>
<td>2</td>
<td>0</td>
<td>0.094240837696353510</td>
<td>0.08376963507853400</td>
<td>0.099476439797057590</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NTHO</td>
<td>2</td>
<td>1</td>
<td>0.1099476497905800</td>
<td>0.1125645026176800</td>
<td>0.08376963507853400</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NTHO</td>
<td>2</td>
<td>2</td>
<td>0.1125645026176800</td>
<td>0.08376963507853400</td>
<td>0.08960282722513090</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CASI</td>
<td>1</td>
<td>0</td>
<td>0.18540268456375800</td>
<td>0.31291946308724800</td>
<td>0.31291946308724800</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CASI</td>
<td>1</td>
<td>1</td>
<td>0.120850369425717500</td>
<td>0.11157714120805400</td>
<td>0.05536912751677850</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CASI</td>
<td>1</td>
<td>2</td>
<td>0.06627516775623490</td>
<td>0.06627516775623490</td>
<td>0.06627516775623490</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CASI</td>
<td>2</td>
<td>0</td>
<td>0.19852979051855000</td>
<td>0.34554973821989590</td>
<td>0.12041884617539000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CASI</td>
<td>2</td>
<td>1</td>
<td>0.053235602094248040</td>
<td>0.49738219895287900</td>
<td>0.0549738219895288</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HHUG</td>
<td>1</td>
<td>0</td>
<td>0.075581395348357200</td>
<td>0.04095767441806490</td>
<td>0.048511279068776700</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HHUG</td>
<td>1</td>
<td>1</td>
<td>0.046511627909676700</td>
<td>0.052325581395348900</td>
<td>0.03779096764418600</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HHUG</td>
<td>2</td>
<td>0</td>
<td>0.10729613739056000</td>
<td>0.17167381974248900</td>
<td>0.14806866952789790</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HHUG</td>
<td>2</td>
<td>1</td>
<td>0.07725321884120200</td>
<td>0.09442060068369100</td>
<td>0.078321680251751070</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HHUG</td>
<td>2</td>
<td>2</td>
<td>0.03755364806666950</td>
<td>0.03755364806666950</td>
<td>0.03755364806666950</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

4. Discussions and Limitations

We have shown that the ETDA framework enables the representation of organized geometric structures, and the topology of cliques that occur during epileptic seizure events, which often occur as abnormal brain activities propagate across different brain regions. The preliminary results demonstrate that the proposed EDTA approach is a promising method to accurately characterize changes brain networks during epileptic seizures.

In this paper, we did not discuss methods to characterize the geometric structures extracted from signal data during clinical events from randomly occurring structures. To address this challenge, we propose to extend previous work by Giusti et al. that discussed the potential use of clique topology to identify a correlation matrix representing functional connectivity metrics that is not random and derived from a geometric matrix (18). Given that a clique topology describes the distribution of order complexes in both random and geometric matrices with the properties of the Betti plots changing with N, we can use the Betti plots to distinguish a random matrix from geometric matrices. We propose to generate random correlation matrices by modifying and shuffling the matrix cell values corresponding to each seizure event in our result with distinct off-diagonal values. Previous studies have reported that the graph in order complexes are nested family of Erdos-Renyi random graph with unimodal Betti curves. The Betti curves of geometric order complexes have distinct characteristics (18). Similarly, we believe that the Betti curves computed from the directed graph models computed from signal data recorded during seizure events will also have distinct properties as compared to random matrices. Using our new persistent homology algorithms for directed graph model, we propose to use the persistence lifetimes to characterize the differences between Betti curves or random and geometric matrices for validating the soundness of the topological structures extracted from the correlation matrices.

5. Conclusions

Neurological disorders affect more than 100 million persons in the nation with significant adverse impact on the affected persons, including disability, morbidity, and economic costs of more than $800 billion per year. Improved understanding of brain connectivity networks that are adversely affected in neurological disorders, such as Epilepsy, Parkinson’s disease, and...
Alzheimer’s disease, can be used to develop more effective therapeutic approaches for patients. However, the interaction dynamics of brain networks in neurological disorders is poorly understood. Although there is increasingly availability of multi-modal data, there are critical limitations of existing graph models used in research studies as graphs are inherently binary representations of complex polyadic or n-ary interactions involving multiple brain regions simultaneously. The goal of this project is to develop new algebraic topology models and algorithms to characterize multi-dimensional interaction structures formed during events related to neurological disorders, such as epileptic seizures with multiple simultaneous foci of interactions. To validate the new algebraic topology structures in neurological disorders, we used high resolution stereotactic electroencephalogram (SEEG) data recorded from a cohort of five patients with refractory epilepsy. In particular, we used clique structured formed during seizure events to compute topological structured and using persistent homology we analyzed the changes in interaction between brain regions during seizure events. Using the metric of bottleneck distance, we compared the differences between persistent diagrams computed for the five patients in this study. The proposed ETDA approach showed promise in analyzing high-dimensional interactions between brain regions during seizure events.

Acknowledgements

This work is supported in part by the National Institutes of Biomedical Imaging and Bioengineering (NIBIB) Big Data to Knowledge (BD2K) grant (1U01EB020955), NSF grant# 1636850.

References


1253
User Needs and Challenges in Information Sharing between Pre-Hospital and Hospital Emergency Care Providers

Zhan Zhang, PhD¹, Aleksandra Sarcevic, PhD², Karen Joy, BS¹, Mustafa Ozkaynak, PhD³, Kathleen Adelgais, MD, MPH³
¹Pace University, New York, NY, USA; ²Drexel University, Philadelphia, PA, USA; ³University of Colorado, Aurora, CO, USA

Abstract

Effective communication between pre-hospital and hospital providers is a critical first step towards ensuring efficient patient care. Despite many efforts in improving the communication process, inefficiencies persist. It is critical to understand user needs, work practices, and existing barriers to inform technology design for supporting pre-hospital communication. However, existing research examining the ways in which patient information is collected and shared by pre-hospital providers in the field has been limited. We conducted a series of ethnographic studies with both pre-hospital and hospital care providers to examine 1) the types of information that are commonly collected and shared by the pre-hospital providers in the field; 2) the types of pre-hospital information that are needed by hospital-based teams for ensuring appropriate preparation; and 3) the challenges in the pre-hospital communication process. We conclude by discussing technology opportunities for facilitating real-time information sharing in the field.

Introduction

In high-risk, time-sensitive medical domains, such as emergency care, medical professionals must provide rapid treatment and manage potentially life-threatening illnesses or injuries (e.g., trauma injuries, stroke, medication overdose). Effective and timely information sharing between pre-hospital and hospital providers (also known as pre-hospital communication) is a critical first step for achieving this goal¹. Information collected in the field and en route to the hospital (pre-hospital information) can help the emergency care providers at the receiving hospital anticipate the severity of patient illness or injury, and make appropriate preparations and triage decisions². Despite its critical role, information sharing between the field providers and those at the receiving hospital remains challenging³, ⁴. For example, verbal reports given by pre-hospital providers during patient transport often lack detail⁵, ⁶ or accuracy⁷, ⁸, making it difficult for hospital teams to appropriately prepare. The highly dynamic nature of out-of-hospital encounters is also characterized by frequent interruptions, posing challenges on real-time data collection and leading to delayed and incomplete information dissemination from the field⁹.¹⁰.

Previous research has developed information and communication technologies (ICTs) to support data transfer from the field to receiving hospitals¹¹, ¹². Key examples include mobile electronic documentation systems¹⁰, ¹³ and ambulance-based telemedicine systems¹⁴⁻¹⁹. These systems, however, are rarely used in real-time due to portability and usability issues¹⁵, ¹⁹, ²⁰. Pre-hospital providers have to perform hands-on examinations and treatments on patients while managing information from multiple sources in short time periods. This work practice limits their direct interaction with handheld systems²¹. Prior research has highlighted that the development of ICTs for healthcare professionals should not only focus on technological aspects but also account for user needs and current work practices³, ¹². Although several studies have looked at information handover workflow between pre-hospital and hospital teams in the receiving care centers (e.g., emergency department)³, ⁸, ⁹, ²², limited research exists on the ways in which pre-hospital information is collected and shared in the field.

The long-term goal of our research is to design and develop novel technologies to better support real-time, seamless data sharing between pre-hospital and hospital teams. To achieve this goal, we first need to answer several fundamental research questions (RQs):

RQ1: What types of information are commonly collected and shared by pre-hospital providers in the field?
RQ2: What types of pre-hospital information are needed by hospital teams for ensuring appropriate preparation?
RQ3: What challenges and barriers exist in the pre-hospital communication process?

In this paper, we describe a mixed-methods ethnographic research conducted with both pre-hospital and hospital providers to answer these research questions and inform technological interventions for facilitating the acquisition and dissemination of pre-hospital information in the field.
Background: Pre-Hospital Communication Process

A typical pre-hospital communication process involves multiple geographically distributed and heterogeneous emergency care teams, including Emergency Medical Services (EMS), Emergency Communication and Information Center (ECIC), and Emergency Department (ED) (Figure 1). Depending on the patient needs, other care teams, such as trauma teams, neurology, and cardiology teams may also be activated and involved in the care process. Following an incident, EMS providers (e.g., paramedics and air-ambulance crews) are dispatched to the scene to provide urgent medical care and transport the patient to the nearest or most appropriate care center. EMS teams collect and manage a variety of information about the patient’s status and clinical needs, which inform treatment decisions in the field. Under certain circumstances (e.g., trauma or burn injuries, cardiac arrest, stroke), EMS teams need to notify the receiving hospital about the patient’s status. By protocol, EMS crews should provide a verbal report via radio (also known as pre-arrival notification) to the 9-1-1 communication center, where dispatchers collect the information and relay it to the ECIC teams—the first point of contact at the hospital for crews transporting patients to the hospital. In some cases, EMS providers would choose to contact ECIC or ED directly via phone. Upon receiving the pre-arrival notification, the ECIC staff (e.g., dispatchers or communication specialists) first call the ED charge nurse or physician on call to relay the reported information. If EMS providers request medical advice, ED physicians are added to the EMS-ECIC call to provide guidance and make decisions. If trauma team activation is needed, ECIC staff sends out a brief notification message to trauma team members via pagers. As the trauma team assembles in the resuscitation room, the ED physician relays known information about the patient and works with other trauma team members to prepare for the patient’s arrival. For other critical cases (e.g., stroke or cardiac arrest), care specialists will also be summoned to the ED for consultation, and the receiving teams (e.g., neurology and cardiology teams) will be notified to get ready.

Methods

Data Collection

Data collection occurred in different time periods. Between 2016 and 2017, we conducted semi-structured interviews with three different care teams in an urban pediatric teaching hospital with a Level I trauma center in the east coast region. Participants included six ED physicians, eight ECIC team members, and 16 trauma team members (five emergency medicine physicians, eight senior surgical residents, one surgical fellow, one respiratory therapist, and one nurse practitioner). The interviews focused on their work practices, pre-hospital information needs, and concerns about receiving and using pre-hospital information. The length of interviews varied depending on their availability (ranging from 15 minutes to one hour). This interview study helped us uncover pre-hospital information types that are critical to the work of emergency care professionals at the point of care. We also gained an in-depth understanding of the challenges faced by hospital-based teams in acquiring and using pre-hospital information. These results informed our following studies with EMS teams.

To understand how pre-hospital information is collected and shared by EMS providers, we first reviewed video recordings of 25 simulations performed in an urban fire-based EMS agency in the mountain region. The simulations were conducted for training purposes. Participants in the simulations were paramedics and emergency medical technicians (EMTs) recruited from the EMS agency; all participants were experienced providers and met local and state authority requirements for staffing an Advanced Life Support (ALS) ambulance. Each simulation team consisted of 4–6 members with a designated team leader, carrying out three different scenarios over a period of 6 months in 2019. The scenarios involved a 15-month-old seizure, a 1-month-old with hypoglycemia, and a 4-year-old clonidine ingestion. The simulations were conducted in a mobile simulation laboratory resembling the back of an ambulance,
using high-fidelity patient mannequins. All simulations were captured by three video cameras: 1) the patient’s overhead view, 2) the foot side of the patient, and 3) a zoom-out view of the entire scene.

To augment the findings from video review, we conducted 45-90-minutes long semi-structured interviews with 13 EMS practitioners, 11 of whom are paramedics, and the other two are EMTs. The interviews were conducted via Zoom between 2020 and 2021. The participants were recruited from four hospital-based EMS agencies, which are part of the 9-1-1 system in an urban area in the US Northeast region. Years of experience range from 7 to 30 years, with two participants being EMS directors. The interviews focused on their work experience and backgrounds, job responsibilities, data collection in the field, communication process with physicians, and challenges in sharing data with the receiving hospital.

All interviews were audio-recorded and transcribed for further analysis. The videos were transcribed using excel sheets to provide a linear list of conversations and activities. Both the university and hospital Institutional Review Board (IRB) approved the studies.

Data Analysis

In reviewing the videos, we used an open coding technique to uncover common information types collected in the field and work practices related to data collection and sharing. Two researchers coded the video transcripts. They first reviewed four randomly selected videos to develop a codebook in an iterative manner (e.g., codes and codebook disagreements were discussed through regular meetings). The codebook defines a set of codes related to types of collected and shared information, types of verbal communication (e.g., inquiry, clarification), instances of non-verbal communication (e.g., note taking, gazing, pointing), and artifacts used. The codebook was then used by the researchers to standardize the coding process. The inter-rater reliability between the two researchers was substantial (Cohen’s Kappa coefficient value is 0.7). We also compared the information collected in the field to the information reported to the hospital. Because EMS verbal reports were omitted in 3 simulation sessions, our analysis focused on the remaining 22 simulations. This video review helped answer the first research question (RQ1).

Data from semi-structured interviews were also analyzed using an open coding technique to answer RQ2 and RQ3. The EMS interview analysis focused on the challenges faced by EMS practitioners in communicating patient data to the receiving facility, while the analysis of interviews with hospital teams focused on their needs and concerns in acquiring pre-hospital information from the fields. All the codes generated through the interview analysis were discussed among researchers to determine which codes to keep, merge, or discard. We then used affinity diagrams to generate high-level themes, followed by identifying representative quotes to support the claims.

Results

We report the results in three parts. First, we describe the information types that are collected and shared by EMS providers in the field. Second, we report the needs and challenges of acquiring pre-hospital information by emergency care professionals at the receiving hospital. Finally, we present the challenges and barriers that EMS providers face in communicating patient data to hospital teams.

Types of Information Collected and Shared by EMS providers in the Field

Of 25 simulations, EMS providers verbally reported the information via radio to the receiving hospital in 22 sessions. The analysis of these 22 simulations showed that EMS providers collected and shared about 18 types of information during pre-hospital care. We grouped these 18 information elements into five high-level categories: demographics, mechanism of injury, physical findings, injuries, and treatments (Table 1). Below we discuss each high-level information category in greater detail.

Demographics: Commonly collected demographic information included patient age, gender, name, weight, and medical history. In particular, age and medical history were inquired by EMS providers in all simulations (22/22), while gender and name were asked in 19 and 13 out of 22 sessions respectively. Because the simulations were situated in the context of pediatric emergency care, demographic information was mainly collected through verbal communications with the patient’s parents or guardians. Among these demographic information types, patient age (22/22) was always included in the EMS pre-arrival notification. In contrast, medical history (1/22) and weight (1/8) were rarely shared with the hospital teams.

Mechanism of Injury (MOI): To come up with an appropriate patient management plan, it is critical for EMS providers to know how the patient got injured or what type of incident occurred. EMS providers specifically asked for this information from the patient’s parents or guardians in almost all simulations (20/22): “About 30 minutes ago I heard
him kind of banging around, and then I walked in just a few minutes ago, and then he was just lying there.” Of the 20 simulations where this type of information was collected, EMS crews shared injury mechanism with the receiving hospital in 16 sessions.

Injuries: Because the simulations were medical resuscitation cases, EMS providers only collected injury information in 6 out of 22 sessions. Their focus was primarily on whether the patient presented any signs of trauma (e.g., swelling, bruises, and lacerations). In the simulations, this information was identified either by talking to the patient’s parents or when a certain injury was reported during physical examination. The injury information was reported to the hospital in two cases.

Physical Findings: The information types collected in this category included vital signs (e.g., heart rate, blood pressure, and respiratory rate) (22/22), breathing (21/22), patient neurological status (e.g., alertness, consciousness) (22/22), pulse (8/22), airway status (8/22), symptoms (7/22), and change of status (13/22). Some physical findings were almost always reported in the notification to the hospital, such as vital signs (19/22) and neurological status (16/22). However, other physical findings were rarely shared with the hospital, e.g., such as breathing (11/21), pulse (0/8), airway (0/8), and change of status (2/13).

Treatments: Commonly collected information related to treatments included IV (intravenous) or IO (intraosseous) access and administration (22/22), oxygen (18/22), and treatment outcome (11/22). But only IV/IO administration (20/22) was consistently reported by EMS providers to the hospital.

Our analysis also showed that the collected information was only reported briefly or partially in many cases (Table 1). For example, vital signs information was only fully reported in 7 sessions, but partially reported in 12 sessions. In those partial reporting cases, only one or two physiological data points (e.g., blood pressure) were shared with the

Table 1. Types and frequency of pre-hospital information collected and shared by EMS providers in 22 simulations.

<table>
<thead>
<tr>
<th></th>
<th>Collected in the field through explicit communication and physical examination</th>
<th>Shared with the hospital via radio</th>
<th>Fully Reported</th>
<th>Partially Reported</th>
<th>Not reported</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demographics</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Name</td>
<td>13</td>
<td>0</td>
<td>0</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>22</td>
<td>22</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Gender</td>
<td>19</td>
<td>14</td>
<td>0</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Medical History</td>
<td>22</td>
<td>1</td>
<td>2</td>
<td>19</td>
<td></td>
</tr>
<tr>
<td>Weight</td>
<td>8</td>
<td>1</td>
<td>0</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>Mechanism of Injury</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Incident details</td>
<td>20</td>
<td>9</td>
<td>7</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Physical Findings</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vital Signs</td>
<td>22</td>
<td>7</td>
<td>12</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Pulse</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>Neurological Status</td>
<td>22</td>
<td>14</td>
<td>2</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Breathing</td>
<td>21</td>
<td>5</td>
<td>6</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Airway</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>Symptoms</td>
<td>7</td>
<td>2</td>
<td>0</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Change of Status</td>
<td>13</td>
<td>2</td>
<td>0</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>Injuries</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Signs of trauma</td>
<td>6</td>
<td>2</td>
<td>0</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Treatments</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IV/IO</td>
<td>22</td>
<td>12</td>
<td>8</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Oxygen</td>
<td>18</td>
<td>8</td>
<td>2</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>Outcomes</td>
<td>11</td>
<td>3</td>
<td>0</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>Medication</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>
hospital while other vital signs were excluded from the verbal report. Another example of partial reporting is related to the mechanism of injury, with many details missing from the pre-arrival notification.

During the interviews, we asked EMS practitioners to explain why the information was only partially shared with the receiving hospital. They stated that the purpose of pre-arrival notifications is to “let the hospital know you are coming, not just give them very comprehensive report” [EMS-P1]. Also, they are concerned that the notification receiver (e.g., dispatcher or ED nurse) could get overwhelmed if they report too much information: “I try to keep it super tight because usually if you give too much information, the person who’s taking the call might not remember all of it. So, we just try to keep it to about 30 to 45 seconds if we can” [EMS-P11].

Similar to previous work\(^1\), we also found that EMS providers already follow a relatively stable structure to construct the pre-arrival notification: “The first thing that I share is usually what I’m rolling in—what the diagnosis is that necessitated us calling ahead of time. And then the patient’s demographics, age, gender, their associated medical history, usually any interventions that I’ve performed, and things that are out of the ordinary. And lastly what the ETA [estimated time of arrival] is to the hospital” [EMS-P12].

**Types of Pre-Hospital Information Needed by the Hospital Teams**

The semi-structured interviews with three different hospital-based teams (ECIC, ED, and trauma teams) helped us identify commonly needed pre-hospital information types by emergency care providers at the point of care (Figure 2). Based on these data, findings from the physical examination were considered the most critical type of pre-hospital information. For example, patient neurological status (e.g., loss of consciousness) and vital signs were viewed as critical by most emergency care professionals because this information helped them anticipate the level of patient acuity. Hospital teams also considered mechanism of injury as an important pre-hospital information type because this information helped them “picture” what happened to the patient: “In addition to vital signs, getting an impression about the patient [is also important]. So, impression could be respiratory distress, or motor vehicle crash. I would like to know what happened first” [ED-P5]. Many care providers also wanted to know what treatments were completed en route, and if any medications were administered. As an ECIC staff explained: “So, like, if they have started a [IV] line, and gave morphine, and if they gave other treatments, I would definitely let ED nurse know, so it does not overlap with ED work. In that way, the ED doctor knows at what time they gave morphine […] If they just gave morphine 5 minutes ago, you don’t want to give it once again” [ECIC-P5].

Information needs, however, could also vary across different professions (Figure 2). For example, ED physicians expressed more interest in knowing details about patient injuries (e.g., the type, severity, and location of the injuries) than other teams. We also observed differences in information needs even among the members of the same team. For

---

\(^1\) Similar to previous work, we also found that EMS providers already follow a relatively stable structure to construct the pre-arrival notification: “The first thing that I share is usually what I’m rolling in—what the diagnosis is that necessitated us calling ahead of time. And then the patient’s demographics, age, gender, their associated medical history, usually any interventions that I’ve performed, and things that are out of the ordinary. And lastly what the ETA [estimated time of arrival] is to the hospital” [EMS-P12].
example, several ED physicians expressed the needs to know injury mechanism, however, it is interesting to see that one ED physician had different opinion on the importance of the mechanican: “Mechanism doesn’t matter. Mechanism increases risk about certain injuries, if you use seat belt, if you get hit by a car, but it is still a matter of what. It just increases your risk, but it doesn’t mean any injury. I’ve seen a kid fell from 70 feet from an apartment building, and he is fine: and I’ve seen kids fell from 10 feet, lay on concrete, then he has major head injury. So, mechanism is useless information but it doesn’t really help you” [ED-P4].

Even though the hospital teams may have different opinions on the importance of different types of pre-hospital information, they all expect a well-structured and clearly articulated EMS report: “I want a short, sweet assessment of three criteria that I need. I need to know is there a physiological change in the patient, is there an obvious fracture, and is there a mechanism. They are pretty objective assessments. […] I don’t need that ‘live in the home,’ all that nuanced information. Give me what I need, and I don’t know if EMS really knows what it is we are looking for. […] As you may see, I am rolling my eyes over and over again, I am like just give me the information” [ED-P3].

Our results also showed several concerns and unmet information needs related to receiving the pre-hospital information from EMS providers. First, four ED physicians and trauma team members indicated the desire to receive more contextual information from the accident scene (e.g., photos and videos) to better anticipate the patient’s needs. They considered visual information such as photographs and videos as a helpful addition to the summary of the patient’s status because they visualized augmented the brief EMS reports. However, this information need is not well supported by current communication mechanisms. Second, not all information types considered necessary were available at all times; even when available, information was not always accurate. Third, the level of detail about certain information needed by the hospital-based teams was not clearly established between pre-hospital and hospital teams, leading to some tensions. EMS teams sometimes did not know how important certain types of information about patients were to the hospital teams. It was therefore challenging for them to prioritize these information items in communicating with hospital teams, causing EMS reports to be either too short or overwhelmingly detailed.

**Challenges Faced by EMS Teams in Sharing Information with the Receiving Hospital**

EMS teams are well aware of the importance of communicating accurate and essential patient information with the receiving hospital: “We are sometimes the only eyes and ears for the physician and the hospital. So, the story that we provide may sometimes be the only story that the clinicians at the hospital have to take care of the patient” [EMS-P4]. However, EMS providers face many challenges in accomplishing this important task, including limitations on collecting and reporting patient data in the field, ineffective communication mechanisms, overloaded EMS systems, and intermediary communication links. Below we describe each challenge in greater detail.

**Limitations on Collecting and Reporting Patient Data in the Field.** It is common that EMS providers cannot always obtain sufficient information in the field due to time pressure or the patient status (e.g., unconscious), limiting their ability to report the needed information to the hospital teams. In those situations, they announce their arrival to the hospital teams with only a fraction of the patient information. As one participant explained: “If the patient is unconscious or the patient might be too critical to be able to answer questions, and there might be nobody with the patient, we can’t get their name or any information on demographics. […] We may just guesstimate [guess and estimate] the patient’s age” [EMS-P1]. In other cases, EMS crews may not even have the capability to notify the hospital as their hands are occupied with stabilizing a critically injured or ill patient: “In some critical cases, like if I’m doing CPR to patient, I can’t take out my phone and start calling the charge nurse” [EMS-P10].

**Ineffective Communication Mechanisms.** EMS providers primarily rely on radio or cellular link to communicate with dispatchers and hospital teams. However, the radio signal is often unstable and fails to work in many areas, causing challenges for efficient and accurate pre-hospital communication: “There are dead zones in the city where your signal is poor, so you might not be able to get through in a timely fashion where you might have to move somewhere, either away from the patient or move the patient to somewhere where you have a stronger signal” [EMS-P8]. The COVID-19 pandemic has exacerbated this challenge, as explained by another EMS participant: “What comes to my mind first is that because it’s a different world right now with COVID. We are wearing N95 respirators, and it’s really, really difficult to speak to anyone over the phone, let alone that you have to request orders from that” [EMS-P12].

Due to the system integration issues, the portable radio carried by EMS providers on the scene cannot be used for pre-arrival notification, forcing the EMS crews to use the radio inside the ambulance. Because of this limitation, the EMS provider driving the vehicle is usually the one giving the notification to the hospital. The driver may not necessarily know the most current patient status since they have difficulty communicating with the co-workers in the back of the ambulance: “If you have a real serious patient and you have two paramedics in the back and an EMT who doesn’t
necessarily know what’s going on is driving. And then that EMT has given the notification, they might not get the right information because they don’t know what’s going on or they might forget something” [EMS-P8].

The limitations of the current communication mechanisms also pose significant challenges in communicating contextual information from the field to the receiving hospital, leading to not only time-consuming verbal descriptions but also misinterpretations of the patient status: “I think the biggest issue is the fact that it’s all verbal and they can’t generally see the patient” [EMS-P4]. To work around this issue, a few participants mentioned that they would take a picture of the accident scene for trauma cases to help ED and trauma teams understand the severity of the accident and consider potential internal injuries: “When I showed up to the scene of a really bad accident, I would take pictures of the accident, and then when I got to the hospital, I would show the trauma doctors because they can see if it was a head-on collision, then they know what kind of injuries they’re expecting and what could be wrong with the patient. [...] I find that could be relevant information to doctors and assisting them” [EMS-P11].

Overloaded EMS Systems. An organizational-level challenge is that the EMS system is usually overloaded. The entire system could become extremely busy in peak hours, especially in large cities. For example, EMS providers need to call the emergency communication center to give the pre-arrival notification, but establishing connection with the dispatcher is sometimes challenging: “It’s only one dispatcher. So, if two ambulances are trying to give notification at the same time, one is stuck in limbo. And they might not ever get through because he’s hung up on details on one call. I can’t even talk to him yet about my call and I’m going to get to the hospital before I get through” [EMS-P8].

To address this challenge, EMS practitioners sometimes call the receiving hospital directly using their personal phones: “Because I’m working in a hospital-based EMS system, I have the phone number for the ER, if I have to give a notification to the specific ER that we work at, I will call them on my phone personally, rather than having to go through the dispatcher” [EMS-P12]. Another participant confirmed this “unofficial” yet effective work practice: “There are times where I can’t get over the radio to the dispatcher. So, I do have to call the hospital notification line directly and speak with the nurse in the ED and let them know we’re coming. That’s like more unofficial way of getting a notification. Meaning it’s not technically the correct way to do it, but I think it’s more efficient” [EMS-P9].

EMS practitioners may also need to consult a physician about treatment plans en route or patient destination. Almost all of our EMS participants have experienced difficulties reaching out to ED physicians: “Common problem that has occurred is that there is a long delay to actually speak to the physician sometimes. Sometimes, it can take anywhere between 15 minutes to half an hour, and that can be frustrating. And obviously it delays patient care. [...] What you’re supposed to do is just wait, because otherwise you get in trouble for doing something without speaking to the physician” [EMS-P1].

Intermediary Communication Links. When giving a notification, EMS practitioners have to communicate with a 9-1-1 or ECIC dispatcher who takes the information and relays it to the next recipient in the communication chain. Our results showed several issues in the EMS-dispatcher communication process. First, some information could get miscommunicated or lost during this transition process, as described by an EMS participant: “A lot of information gets lost in translation. I’ve had plenty of times where we show up and the hospital’s ready for a patient in ‘cardiac arrest.’ And we’re like, ‘wait, we did not say that.’ Or we show up and they’re like, ‘we didn’t get a notification about what you’re talking about.’ So, the game of telephone, I think, is a big hindrance in getting notifications. [...] The dispatcher has to understand 100% when you say over the radio and then they have to say it again to a person [ER nurse or physician] on the other end of the phone. And they don’t really have the opportunity to call you back and ask further information. Because as soon as they’re done with that one, there’s another ambulance who’s on the radio waiting to get connected” [EMS-P13].

Second, the dispatchers often asked the EMS practitioners about unnecessary information, which delayed the notification process. An EMS crew member explained: “They [dispatchers] tend to follow a protocol. [...] It’s ingrained in them that they need this and that information or they can’t continue on, even if the information really isn’t going to make a difference. If you call about a stroke patient, they will ask if it is the left side or the right side. It doesn’t matter. Or exact numbers on a set of vitals. [...] If it’s normal, it doesn’t matter. It’s not going to affect the main goal for a notification, which is to have a team ready and the right sources. [...] But they’ll get hung up on those details and that’ll just delay them” [EMS-P8].

Given those challenges, EMS participants expressed their interest in sharing original patient data directly with the receiving ED department: “It’d be great if we are able to send our notes to the hospital, so they could know what’s going on even before we get there” [EMS-P10]. However, this need is not fully supported by the current system architecture and communication mechanisms.
Discussion

We found that EMS providers collected a variety of information from multiple sources in a short time period, including patient demographics, mechanism of injury, physical exam findings, injuries, and treatments. Due to the time pressure, EMS providers were only able to provide a very brief verbal report to the receiving hospital. Our results showed that only a few information types (e.g., patient age, neurological status, vital signs, and IV/IO administration) were always shared by EMS providers, while other types (e.g., details of mechanism of injury, airway, breathing, and change of status) were rarely reported. Even when available, much information, including the essential parts such as vital signs, was often partially reported. A possible reason for this limited or partial reporting may lie in the purpose of the EMS verbal report, which is to quickly announce the patient arrival to get the hospital teams and resources ready. However, these limited or partial reports could also lead to challenges in establishing common understandings between the pre-hospital and hospital teams.

Interviews with different hospital-based teams uncovered information types that are critical to their work. We found that the pre-hospital information needs of hospital teams match what is typically reported by EMS teams. However, hospital teams need contextual information from the accident scene (e.g., photos, videos) to better anticipate the patient’s needs. The challenge, however, is that EMS teams currently lack effective mechanisms by which context-specific information is accrued to allow for rapid integration and sharing. Even in this era of modern communications, EMS providers still rely on radio or phone to communicate with hospital teams. Theses outdated mechanisms make it challenging for EMS providers to not only describe complex patient cases in words but also for hospital teams to understand the symptoms and status of the patient. The limitations of current communication mechanisms pose challenges to efficient information sharing between pre-hospital and hospital teams, requiring further studies.

Despite being aware of the importance of pre-hospital communication, EMS teams face many challenges in delivering accurate information to the receiving hospital in a timely manner. One prominent barrier is related to the limited ability to notify hospitals when dealing with a critical patient in the field, as EMS providers need to perform hands-on examination and treatments on patients. As a result, they often experience high physical and cognitive workload, limiting their abilities to use handheld radio or phone for notification. Another significant challenge is the difficulty of communicating with the dispatchers at the emergency communication centers. Sometimes dispatchers are overloaded, forcing EMS providers to wait in a queue to get connected. This challenge could severely delay the notification process and ultimately patient care. In other cases, miscommunication or information loss could occur in the notification process due to various reasons, such as unstable radio signal or inaccurate interpretation. Because of this issue, the information delivered to ED physicians is not always accurate, hindering their decision-making process (e.g., which specialist to call, what resources and equipment to prepare, and what labs to order).

Given these challenges, it is critical to consider novel technology interventions to support seamless, real-time pre-hospital information sharing in the field. For example, Schooley et al. reported the design and evaluation of an Android-based smartphone application to support the communication of patient information from the field to the receiving hospital. EMS providers can use this application to take photographs, record digital audio notes, and capture video of the patient and scene. Such data can be relayed to and reviewed by ED physicians via a web application. An ambulance-based telemedicine system is another key example of technology being tested over the past decade to enable real-time, audio-video pre-hospital communication. The system is installed in the ambulance with video cameras to capture a designated area’s view (e.g., patient body) and transfer data to the hospital. Despite the benefits, various challenges hinder the effective use of these technologies. For example, given the size and weight of the ambulance-based telemedicine unit, this system is not portable enough to be used outside of the ambulance where a great portion of urgent patient care occurs. As a result, these handheld systems have rarely been used in real time, becoming a hindrance to the user.

Recent work has shown that “smart glasses” have a high potential for supporting pre-hospital information transfer, because this novel technology offers touchless interaction mechanisms, such as voice control, to minimize the intrusiveness of these tools. These head-mounted, wearable devices with a transparent screen and a video camera that can project first person, point-of-view data to a remote viewer are hands-free, allowing care providers to focus on patients in many hands- and eyes-busy medical contexts. This novel technology could also reduce the likelihood of cross-contamination and patient infections since care providers do not have to handle the device physically.

Smart glasses also allow EMS providers to capture and share videos and pictures in real time, which are considered useful by emergency care professionals in the hospital. By doing so, EMS providers can easily share specific information about a patient (e.g., injury location and severity, symptoms) instead of spending a significant amount of time describing the patient situation with words. In addition, pre-hospital and hospital providers can connect via...
videoconferencing applications within the smart glass for more effective care coordination. In this case, ED physicians can see and hear what the EMS providers see and hear through the smart glass application, allowing them to offer medical advice as EMS crews manage and stabilize a severe patient during patient transport. All these promising features of smart glasses make this technology of interest to EMS professionals, allowing them to have faster access to expert advice anywhere (e.g., outside of the ambulance). In our future work, we will look into how to design and develop smart glasses and hands-free interaction mechanisms to support real-time and effective pre-hospital information collection and sharing in the field. As prior work pointed out, smart glasses may introduce cognitive burden and human factor issues. These limitations require considering how smart glasses could be designed for seamless integration into the workflow of pre-hospital care while accounting for physical and cognitive limitations of emergency care providers. Also, social, organizational, and policy factors that can facilitate or impede the use and uptake of smart glasses will be explored in our future studies with various stakeholders.

This study has several limitations. First, we mainly relied on video review of simulated EMS interventions to understand the types of pre-hospital information being collected and shared in the field. It is possible that video recordings of simulations may not have clearly captured all work practices. Despite this limitation, this video review allowed us to analyze the data offline and capture detailed communication and work practices. Also, simulations provide a safe environment without the risk of loss of patient confidentiality. Second, the video recordings we analyzed were simulations with only 3 medical resuscitation scenarios. This limitation might affect the generalizability of the results to other domains (e.g., stroke or trauma). In our future work, we will conduct additional field studies with EMS providers to cover as many patient scenarios as possible for a more comprehensive understanding of the pre-hospital data collection and sharing practices in real-world situations. Third, there is a five-year gap between the interviews with hospital teams and EMS teams. However, since the work practices and communication technologies (e.g., radio signal) haven’t changed during this time period, we believe the impact of this gap on our findings is limited. Lastly, the video data and interview data were collected from different EMS agencies in different regions. However, these EMS agencies follow similar protocol and regulation, making the results generalizable. In addition, their different characteristics (e.g., fire-based vs. hospital-based EMS agency, east coast region vs. mountain region) can further strengthen the generalizability of the results.

Conclusion

In this study, we conducted a series of mixed-methods ethnographic studies with both pre-hospital and hospital care providers to examine user needs, work practices, and existing barriers in the pre-hospital communication process. We described the types of pre-hospital information commonly collected and shared in the field and whether the shared data meet the information needs of emergency care providers in the receiving hospital. We also discussed the issues and challenges related to real-time information sharing during pre-hospital encounters from both pre-hospital and hospital teams’ perspectives. Finally, we used these findings to discuss potential technology solutions that could address the identified challenges and support seamless information sharing between pre-hospital and hospital teams.

Acknowledgement

This work was supported by National Science Foundation (NSF) Award #1948292 and #1253285. We thank Randall S. Burd, MD, PhD; Jennifer Fritzeen, MSN, RN, TCRN, CCNS; Michael Courtney; and, Jack Finkelstein for their support. We also thank EMS practitioners, ECIC staff, and trauma team members for their participation in this study.

References

Data and Model Biases in Social Media Analyses: A Case Study of COVID-19 Tweets

Yunpeng Zhao, MS1, Pengfei Yin, MS1, Yongqiu Li, BS1, Xing He, MS1, Jingcheng Du, PhD2, Cui Tao, PhD2, Yi Guo, PhD1, Mattia Prosperi, PhD1, Pierangelo Veltri, PhD3, Xi Yang, PhD1, Yonghui Wu, PhD1, Jiang Bian, PhD1.

1University of Florida, Gainesville, Florida, USA; 2University of Texas Health Science Center at Houston, Houston, Texas, USA; 3Magna Graecia University of Catanzaro, Catanzaro, Italy

Abstract

During the coronavirus disease pandemic (COVID-19), social media platforms such as Twitter have become a venue for individuals, health professionals, and government agencies to share COVID-19 information. Twitter has been a popular source of data for researchers, especially for public health studies. However, the use of Twitter data for research also has drawbacks and barriers. Biases appear everywhere from data collection methods to modeling approaches, and those biases have not been systematically assessed. In this study, we examined six different data collection methods and three different machine learning (ML) models—commonly used in social media analysis—to assess data collection bias and measure ML models’ sensitivity to data collection bias. We showed that (1) publicly available Twitter data collection endpoints with appropriate strategies can collect data that is reasonably representative of the Twitter universe; and (2) careful examinations of ML models’ sensitivity to data collection bias are critical.

Introduction

The coronavirus disease (COVID-19) pandemic has put tremendous strain on the society. As of March 9, 2021, more than 29.1 million Americans have been diagnosed with COVID-19 and more than 526,000 have died.1 Governments worldwide are trying their best to contain the spread of the virus. Preventative measures, such as social distancing, school closures, and work-from-home policies, implemented by national, state, and local governments, have affected the daily routines of billions of people worldwide and forced many activities and social interactions to be moved online.2,3 Social media platforms are a way for people stay connected during this pandemic. Individuals are increasingly sharing a large amount of personal health information, including their COVID-19-related sentiments and comments. Officials such as health organizations and government agencies have used social media to share COVID-related policies, progress of vaccine development, and Q&A towards COVID-19 issues to help the public stay safe and informed.4 These social media data provide unique insights into public health events.

Among the popular social media platforms, Twitter, initially a microblogging platform, has well-constructed Application Programming Interfaces (APIs) for obtaining the data that are publicly available. Therefore, it has become a popular source of social media data for researchers. In the short time since the pandemic began, Twitter has been used to study various topics around COVID-19. For example, Kouzy et al. (2020) manually identified and quantified the magnitude of misinformation that is being spread on Twitter regarding the COVID-19 pandemic which can served as an early warning for unexpected events.5 Xue et al. (2020) used latent Dirichlet allocation (LDA) to identify popular topics and sentiments from 4 million COVID-19 tweets.6 Mackey et al. (2020) used the Biterm topic modeling to identify individual reports of COVID-19-related symptoms, testing, and recoveries that appeared on Twitter.7

Twitter provides new opportunities for health-related research. However, the use of Twitter data for research also has drawbacks and difficulties; the potential for bias appears at every stage of the process, from data collection to modeling. Tools for collecting social media data often result in biased samples. For example, Twitter APIs only return a subset of the tweets from Twitter’s data warehouse; but the relative size of a subset as a proportion of the whole is unknown, as are the sampling strategies used by Twitter to produce the subsets.8 Further, social media data (e.g., tweets) are mostly unstructured free-text data. To study health information on Twitter, researchers often use inference models such as machine learning (ML) and topic modeling methods to process and identify insights from these free-text tweets. However, supervised ML models (e.g., models used to identify genuine laypeople discussions from health-related discussions on Twitter9–11) require annotated training samples. If ML models are trained on

---

a Corresponding: Jiang Bian, PhD; bianjiang@ufl.edu
potentially biased subsets of data, it is not clear how well those models will perform when analyzing other samples from the whole datasets. Biases of social media data and analysis methods have yet to be rigorously addressed, particularly in public health surveillance studies such as those for COVID-19. Overlooking these biases in health-related social media studies can lead to wrong or inappropriate results with severe unintended consequences.

Thus, in this study, we used six different data collection methods available to collect COVID-19-related Twitter data and developed three machine learning models. This study had two primary aims, listed below with three corresponding research questions (RQs):

Aim 1: Assess data collection bias.
- RQ1: What proportions of each dataset data collection method returns?
- RQ2: How representative of the data collected with each of the 5 data collection methods to the gold standard dataset (i.e., from the Twitter full archive)?

Aim 2: Measure ML models’ sensitivities to data collection bias.
- RQ3: How does data collection bias (i.e., models trained on different subsets) impact models’ performance when applied on other subsets?

Methods

Figure 1 shows the overall process of our study, where we (1) collected COVID-19 related tweets using 6 different data collection methods, (2) estimated the data collection biases in two-fold: i) assessing the overlapping portions between each pair of the six datasets, and ii) assessing the data representativeness using rank correlation based on the top keywords comparing each of the other 5 data collection methods against the full archive data, and (3) explored ML models’ sensitivity to data bias in terms of model performance by training and testing prediction models on samples selected from different datasets.

Data collection

From February 21st, 2020 to May 1st, 2020, we collected COVID-19-related tweets using three different Twitter APIs: (1) Twitter search API (i.e. “GET search/tweets”), (2) Twitter sampled stream API (i.e. “GET statuses/sample”), and (3) Twitter filtered stream API (i.e. “POST statuses/filter”) using a list of keywords (e.g. “#coronavirus” and “covid”). The list of keywords was developed through a snowball sampling process, where we started with a list of seed keywords collected from online information sources such as news sites and Wikipedia. We then iteratively queried sample tweets from the Twitter website using these keywords and manually reviewed the content of the tweets to discover new COVID-19-related keywords (i.e., words that co-occur with one of the existing keywords but that were not in the existing keyword list) until no new keywords were found. Through this process, we initially found 36 COVID-19-related keywords initially.

As the COVID-19 pandemic progressed, using a similar process, we further extended our 36 keywords to a total of 86 keywords based on a new round of sampling of relevant tweets and Google search results. Since Twitter APIs have rate limits (e.g. for the search AP, only 450 requests are allowed in a 15-minute time window for each crawler and restricted by IP), we split the 86 keywords into 4 groups and used 4 crawlers making 4 sets of Twitter search API requests separately (a.k.a “Twitter multi-search” in our experiments) from May 14, 2020 to July 17, 2020.

On April 29, 2020, Twitter released a new streaming interface that is designed for COVID-19 data collection (a.k.a a “Twitter COVID-19 stream” in our experiment). This COVID-19 endpoints allow approved developers to access COVID-19 related tweets across languages. Nevertheless, the exact mechanism of how Twitter decides that a tweet related to COVID-19 is unclear. Thus, to make this COVID-19 streaming dataset comparable to our other Twitter...
data collection methods, we filtered the dataset using the same set of keywords and the same time range as in our multi-crawler data collection method.

In January 2021, Twitter made its full archive data available to academic researchers; however, the full archive APIs are still constrained by the rate limits.16,17 Ideally, the full archive dataset should contain all tweets of the Twitter universe and thus can be used as the gold-standard dataset for calculating the data coverages of the different data collection methods described above. Because of the rate limits, we collected 200 random samples of a full-minute of tweets from the “Twitter full archive” based on the corresponding keywords and the time ranges that we used for the other data collection methods accordingly.

In summary, we used six Twitter data collection methods resulting in six different datasets: (1) "Twitter search", (2) "Twitter filtered stream", (3) "Twitter sampled stream", (4) "Twitter COVID-19 stream", (5) "Twitter multi-crawler search", and (6) "Twitter full archive". Figure 2 (A) shows the time range of each dataset; and as shown in Figure 2 (B), we divided the datasets into two groups based on the data collection time ranges. Group 1 includes four datasets (i.e. "Twitter search", "Twitter filtered stream", "Twitter sampled stream", and "Twitter full archive") collected from February 21, 2020 to May 1, 2020; and Group 2 includes three datasets (i.e. "Twitter COVID-19 stream", "Twitter multi-crawler search", and "Twitter full archive") collected from May 14, 2020 to July 17, 2020. As shown in Figure 2 (B), ideally, each of the other Twitter data collection methods returns a subset of the "Twitter full archive" dataset, while overlaps with the other datasets collected from the same time period and with the same list of keywords. However, since we collected the full archive data in January 2021, tweets that have been deleted or from accounts that have been suspended were no longer accessible, resulting in other datasets having tweets not in the full archive. The eight circles in Figure 2 (B) represent eight random samples that we used to train our ML models. The results of the training sample annotations and associated ML model performance will be detailed in the result section.

Bias in social media data collection methods
The underlying mechanisms of the Twitter APIs’ sample selection strategies are unknown. We first assessed the data collection bias in terms of overlaps among the datasets within the two groups (as shown in Figure 2 (B)). We then compared the overlapping of tweets by considering the top relevant keywords and measuring the Kendall correlations for each dataset against the full archive dataset.

Originally, we thought that the Twitter full archive API could reliably be used as the gold-standard dataset; however, since we collected the full archive data in January 2021, tweets that had been deleted and tweets from suspended Twitter accounts were no longer available. Thus, when we compared each dataset with the full archive, these tweets had to be removed from our calculations.

In addition to comparing the first five data collection methods with the "Twitter full archive" dataset over the span of a minute, we also measured the overlapping among the first five datasets within additional time spans, including a minute, an hour, a day, and a week. To do so, we first randomly selected 100 random time periods at the lengths of a minute, hour, day, and week. We then calculated the overlapping proportions of the datasets generated from each data collection method and reported the mean overlapping proportions and associated confidence intervals across 100 random samples. For Group 1, we calculated the overlapping proportions among the datasets using the total of the first three datasets (i.e., “Twitter search”, “Twitter filtered stream”, and “Twitter sampled stream”); and for Group 2, the denominator consisted of the total of the “Twitter multi-crawler search” and “Twitter COVID-19 stream” datasets. In addition to calculating the global overlapping proportions among these five datasets, we also aimed to answer RQ2:
“how representative of the data collected with each of the 5 data collection methods to the gold standard dataset (i.e. from the Twitter full archive)?”. To make a fair comparison, we first identified the top 10 keywords of each dataset. We then measured the overlapping proportions between each of the first five datasets and the gold-standard dataset, only considering tweets containing the top 10 keywords. To measure the representativeness, we performed Kendall correlations for each comparison.

**Machine learning models’ sensitivity to data bias**

To test machine learning modes’ sensitivity to data bias, we considered a prediction task that classifies task, based on our past work,\(^9\) that classifiers each tweet into two categories: promotional information and consumer discussions.

**Training sample selection.** We randomly selected 600 English tweets from eight different sampling points (4,800 total tweets) as shown in Figure 2 (B). Those 4,800 tweets were manually sorted into the 3 groups (i.e., irrelevant, promotional, and consumer discussions) by three annotators (kappa = 0.73). Even if a tweet contains keywords related to COVID-19, the tweet may not be relevant or meaningful (e.g., “#SIRC$WORL$DNEWS #Coronavirus #CoronavirusOutbreak -- BREAKING - <URL>”); thus, we categorized those tweets as irrelevant. Within the relevant tweets, we further categorized those tweets into promotional information (e.g., from a health organization “CDC Denies Delaying Testing of California Coronavirus Patient; Over 100 CA Hospital Workers in Home Quarantine <URL>”) and consumer discussions (e.g., “My cousin just died of Coronavirus <URL>”).

**Tweet text preprocessing.** To build the classifiers, we first preprocessed the sampled tweets following the preprocessing steps used by GloVe:\(^8\) (1) removed hyperlinks, (2) removed mentions, (3) replaced hashtags into English words with hashtag sign (e.g., convert “#COVID” to “<hashtag> COVID”), and (4) replaced all emojis, URLs, and mentions (e.g., @username) with signs of “<emojis>,” “<url>,” and “<user>” respectively.

**Machine learning model sensitivity to data bias.** We explored three commonly used classification algorithms: convolutional neural networks (CNN), random forest (RF), and gradient boosting trees (GBT).\(^{19-21}\) We implemented the CNN models in Keras on top of the Tensorflow framework. We initialized the embedding layer with the GloVe pre-trained 100-dimension Twitter word embeddings. We implemented the RF and GBT via the scikit-learn library and used the Term Frequency-Inverse Document Frequency (TF-IDF) scheme to convert each tweet into a feature vector.

Regarding our research question about “model sensitivity to data collection bias”, our hypotheses were: (1) prediction models that were trained on samples selected from one dataset may not achieve consistent performance when they are applied on data from other datasets; and (2) prediction models trained on samples selected from the “Twitter full archive” dataset should achieve relatively higher performance and higher consistency compared to the models trained by samples from the other collection methods, since these are theoretically samples of the full archive.

To test our hypotheses, we trained each model 10 times on each of the eight annotated samples and then tested the model’s performance on the other samples. The performances were measured in terms of mean F-1 score and the 95% confidence intervals (CI) were reported across 10 runs.

**Results**

**Data collection**

In total, we collected more than 750 million tweets using six different Twitter APIs across two different time periods.

<table>
<thead>
<tr>
<th>Data collection methods</th>
<th>Number of tweets</th>
<th>Time range</th>
<th>Number of keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>Twitter search API</td>
<td>200,423,651</td>
<td>02/21/20 - 05/01/20</td>
<td>36</td>
</tr>
<tr>
<td>Twitter filtered stream API</td>
<td>108,987,452</td>
<td>02/21/20 - 05/01/20</td>
<td>36</td>
</tr>
<tr>
<td>Twitter sampled stream API</td>
<td>3,145,428</td>
<td>02/21/20 - 05/01/20</td>
<td>36</td>
</tr>
<tr>
<td>Twitter multi-crawler search</td>
<td>253,996,071</td>
<td>05/14/20 - 07/17/20</td>
<td>86</td>
</tr>
<tr>
<td>Twitter COVID-19 stream</td>
<td>183,815,527</td>
<td>05/14/20 - 07/17/20</td>
<td>86</td>
</tr>
<tr>
<td>Twitter full archive</td>
<td>298,040</td>
<td>02/21/20 - 05/01/20</td>
<td>36</td>
</tr>
<tr>
<td></td>
<td>280,342</td>
<td>05/14/20 - 07/17/20</td>
<td>86</td>
</tr>
</tbody>
</table>

**Data collection bias analysis**

**RQ1: What proportions of each dataset data collection method returns?**

We randomly selected 100 random time periods at the minute, hour, day, and week scales within each of the two datasets. We first compared each dataset against the gold standard dataset at the minute scale by calculating
the overlapping proportion between each dataset and the “Twitter full archive” dataset. The denominator used for this comparison was the total number of tweets in the “Twitter full archive” dataset. We then compared the overlapping portions of the datasets within each group, using the combination of all available datasets within each group as the denominator. We also identified the top 10 keywords of each dataset as shown in Table 2. Except the “Twitter full archive” data in Group 1, the other datasets collected at the same time period have the same top 10 keywords within its group. The “Twitter full archive” data in Group 1 has 9 keywords that are the same as the other three datasets within that same group, with the exception of "#viruscorona" which only exists in the top 10 list of the other sample datasets.

Table 2. Top 10 keywords and corresponding number of tweets by datasets.

<table>
<thead>
<tr>
<th>Group</th>
<th>Datasets</th>
<th>Top 10 keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Twitter search</td>
<td>coronavirus (n=72,165,676), covid-19 (n=24,390,634), #coronavirus (n=21,836,606), novel coronavirus (n=550,429), #covid-19 (n=503,500), ncov (n=315,694), #coronavirusoutbreak (n=243,728), #wuhanvirus (n=220,353), 19-ncov (n=64,158), #viruscorona (n=60,339)</td>
</tr>
<tr>
<td></td>
<td>Twitter filtered stream</td>
<td>coronavirus (n=77,382,453), #coronavirus (n=22,722,083), covid-19 (n=21,630,298), #covid-19 (n=524,053), novel coronavirus (n=455,227), #coronavirusoutbreak (n=383,747), ncov (n=376,125), #viruscorona (n=134,011), #wuhanvirus (n=85,157), 19-ncov (n=49,960)</td>
</tr>
<tr>
<td></td>
<td>Twitter sampled stream</td>
<td>coronavirus (n=1,273,468), covid-19 (n=434,379), #coronavirus (n=352,080), ncov (n=13,598), #covid-19 (n=8,477), #coronavirusoutbreak (n=6,133), novel coronavirus (n=5,436), #wuhanvirus (n=4,322), #viruscorona (n=1,257), 19-ncov (n=430)</td>
</tr>
<tr>
<td></td>
<td>Twitter full archive</td>
<td>coronavirus (n=188,875), covid-19 (n=75,789), #coronavirus (n=62,469), #covid-19 (n=1,919), novel coronavirus (n=1,848), #coronavirusoutbreak (n=1,076), ncov (n=883), 19-ncov (n=321), 2019-ncov (n=319), #wuhanvirus (n=273)</td>
</tr>
</tbody>
</table>

Table 3 (considering all the tweets in each dataset) and Table 4 (considering tweets that contain the top 10 keywords) show the mean number of tweets and 95% confidence intervals across the 100 random samples at each time scale. The "Twitter COVID-19 stream" and "Twitter multi-crawler" search APIs collected higher volumes of tweets compared with "Twitter search", "Twitter filtered stream", and "Twitter sampled stream" APIs at each time scale.

Table 3. Tweet counts including all the tweets in each dataset at each time scale.

<table>
<thead>
<tr>
<th>Scale</th>
<th>All the tweets of each dataset</th>
<th>Group 1 (02/21/20 - 05/01/20)</th>
<th>Group 2 (05/14/20 - 07/17/20)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Twitter search</td>
<td>Twitter filtered stream</td>
<td>Twitter sampled stream</td>
</tr>
<tr>
<td></td>
<td>Minute (Overlapping with Archive)</td>
<td>1,325</td>
<td>1,232</td>
</tr>
<tr>
<td></td>
<td>1,132</td>
<td>1,192</td>
<td>1,272</td>
</tr>
<tr>
<td></td>
<td>1,518</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Minute</td>
<td>2,416</td>
<td>1,854</td>
</tr>
<tr>
<td></td>
<td>183</td>
<td>2,220</td>
<td>1,191</td>
</tr>
<tr>
<td></td>
<td>5,118</td>
<td>52,420</td>
<td>100,777</td>
</tr>
<tr>
<td></td>
<td>Hour</td>
<td>142,004</td>
<td>82,779</td>
</tr>
<tr>
<td></td>
<td>12,894</td>
<td>52,420</td>
<td>100,777</td>
</tr>
<tr>
<td></td>
<td>286,153</td>
<td>1,272</td>
<td>2,205,746</td>
</tr>
<tr>
<td></td>
<td>Day</td>
<td>3007246</td>
<td>1,869</td>
</tr>
<tr>
<td></td>
<td>(38,773, 5,030,460)</td>
<td>1,272</td>
<td>2,205,746</td>
</tr>
<tr>
<td></td>
<td>11503525, 26142006</td>
<td>1,272</td>
<td>2,205,746</td>
</tr>
<tr>
<td></td>
<td>Week</td>
<td>18,844,663</td>
<td>15142006</td>
</tr>
<tr>
<td></td>
<td>18,138,284, 26,844,663</td>
<td>11503525, 26142006</td>
<td>161826,306, 560712</td>
</tr>
</tbody>
</table>
Table 4. Tweet counts and 95% confidence intervals across the 100 random samples at each time scale.

<table>
<thead>
<tr>
<th>Scale</th>
<th>Twitter search</th>
<th>Twitter filtered stream</th>
<th>Twitter sampled stream</th>
<th>Twitter full archive</th>
<th>Twitter COVID-19 stream</th>
<th>Twitter multi-crawler search</th>
<th>Twitter full archive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tweets that contain the top 10 keywords</td>
<td>Group 1 (02/21/20 - 05/01/20)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Minute (Overlapping with Archive)</td>
<td>884 (741, 1,027)</td>
<td>943 (912, 973)</td>
<td>18 (16, 20)</td>
<td>1,856 (1,718, 1,994)</td>
<td>705 (558, 656)</td>
<td>902 (842, 962)</td>
<td>915 (856, 974)</td>
</tr>
<tr>
<td>Hour</td>
<td>1,753 (145, 4,249)</td>
<td>1,674 (856, 2,112)</td>
<td>29 (7, 53)</td>
<td>NA</td>
<td>2,386 (1,656, 3,116)</td>
<td>2,666 (2,496, 2,835)</td>
<td>NA</td>
</tr>
<tr>
<td>Day</td>
<td>59,906 (55,148, 64,664)</td>
<td>50,389 (42,257, 58,521)</td>
<td>1120 (980, 1,259)</td>
<td>59,906 (55,148, 64,664)</td>
<td>143,564 (140,646, 146,483)</td>
<td>162,299 (159,636, 164,962)</td>
<td>NA</td>
</tr>
<tr>
<td>Week</td>
<td>2,413,103 (29,506, 4,372,090)</td>
<td>1,714,022 (1,173,389, 2,081,265)</td>
<td>50,407 (43,213, 57,601)</td>
<td>NA</td>
<td>3,405,842 (3,156,694, 3,654,991)</td>
<td>3,850,300 (3,665,628, 4,034,972)</td>
<td>NA</td>
</tr>
</tbody>
</table>

Table 5 shows the overlapping proportions of each pair of datasets from Group 1. As shown in Table 5, the "Twitter search" dataset captured most of the relevant tweets (i.e., a larger portion than the others). The overlapping between the "Twitter search" dataset and the "Twitter filtered stream" are slightly increased at all time scales, comparing only the tweets with the top 10 keywords.

Table 5. Overlapping data among datasets in Group 1.

<table>
<thead>
<tr>
<th>Scale</th>
<th>Twitter search</th>
<th>Twitter filtered stream</th>
<th>Twitter sampled stream</th>
<th>Twitter search vs. Twitter filtered stream</th>
<th>Twitter full archive</th>
<th>Twitter COVID-19 stream</th>
<th>Twitter multi-crawler search</th>
<th>Overlapping across the three datasets</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(95% confidence intervals)</td>
</tr>
<tr>
<td>Minute</td>
<td>52.3% (45.1%, 59.6%)</td>
<td>48.2% (44.7%, 51.6%)</td>
<td>0.8% (0.7%, 0.8%)</td>
<td>28.6% (23.4%, 33.3%)</td>
<td>0.4% (0.3%, 0.5%)</td>
<td>0.6% (0.6%, 0.7%)</td>
<td>0.4% (0.3%, 0.4%)</td>
<td></td>
</tr>
<tr>
<td>Minute</td>
<td>72.7% (49.7%, 95.7%)</td>
<td>64.5% (31.8%, 97.2%)</td>
<td>1.6% (1.3%, 1.9%)</td>
<td>37.7% (23.1%, 52.4%)</td>
<td>0.6% (0.2%, 1.0%)</td>
<td>0.6% (0.5%, 0.8%)</td>
<td>0.6% (0.1%, 1.0%)</td>
<td></td>
</tr>
<tr>
<td>Hour</td>
<td>72.9% (51.2%, 94.6%)</td>
<td>54.7% (17.8, 91.6%)</td>
<td>1.8% (0.9%, 2.7%)</td>
<td>28.4% (9.0%, 47.8%)</td>
<td>0.5% (0.3%, 0.7%)</td>
<td>0.6% (0.4%, 0.7%)</td>
<td>0.4% (0.1%, 0.7%)</td>
<td></td>
</tr>
<tr>
<td>Day</td>
<td>70.5% (60.1%, 80.9%)</td>
<td>53.4% (28.2%, 78.6%)</td>
<td>1.9% (1.0%, 2.9%)</td>
<td>25.6% (8.4%, 42.8%)</td>
<td>0.5% (0.3%, 0.7%)</td>
<td>0.7% (0.5%, 0.7%)</td>
<td>0.4% (0.1%, 0.7%)</td>
<td></td>
</tr>
<tr>
<td>Week</td>
<td>70.4% (65.6%, 75.2%)</td>
<td>54.0% (37.6, 70.4%)</td>
<td>1.9% (1.3%, 2.5%)</td>
<td>27.5% (16%, 39%)</td>
<td>0.5% (0.2%, 0.7%)</td>
<td>0.6% (0.4%, 0.7%)</td>
<td>0.4% (0.2%, 0.5%)</td>
<td></td>
</tr>
</tbody>
</table>

Table 6 shows the overlapping proportions between each pair of the datasets in Group 2. As shown in Table 6, "Twitter multi-crawler search" covers a higher proportion than "Twitter COVID-19 stream". The overlapping between "Twitter multi-crawler search" and "Twitter COVID-19 stream" is over 50% at all time scales. When comparing only the tweets containing the top 10 keywords, the overlapping between these two datasets increased to almost 60%.

Another interesting finding is that the "Twitter multi-crawler search" covered 98.4% of the tweets with the top 10 keywords, which shows that using multiple Twitter search crawlers and multiple queries can increase the data coverage.
Table 6. Data overlapping among datasets in Group 2.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Minute(^a)</td>
<td>62.8% (61.9%, 63.8%)</td>
<td>76.0% (74.2%, 77.8%)</td>
<td>51.8% (50.4%, 53.2%)</td>
<td>78.0% (74.9%, 81.1%)</td>
<td>98.4% (97.7%, 99.1%)</td>
<td>77.3% (74.2%, 80.4%)</td>
</tr>
<tr>
<td>Minute(^b)</td>
<td>73.7% (72.8%, 74.6%)</td>
<td>78.9% (77.4%, 80.5%)</td>
<td>52.6% (58.8%, 61.8%)</td>
<td>75.8% (74.0%, 77.6%)</td>
<td>84.5% (83.8%, 85.2%)</td>
<td>60.3% (58.8%, 61.8%)</td>
</tr>
<tr>
<td>Hour(^b)</td>
<td>67.1% (51.3%, 82.7%)</td>
<td>82.3% (68.5%, 95.5%)</td>
<td>50.5% (30.1%, 62.9%)</td>
<td>74.52% (73.5%, 75.5%)</td>
<td>84.1% (83.7%, 84.5%)</td>
<td>58.6% (57.8%, 59.4%)</td>
</tr>
<tr>
<td>Day(^b)</td>
<td>69.2% (57.3%, 80.8%)</td>
<td>83.1% (73.4%, 92.6%)</td>
<td>52.1% (43.4, 60.6%)</td>
<td>74.42% (74.2%, 74.7%)</td>
<td>84.5% (84.3%, 84.9%)</td>
<td>59.8% (58.2%, 59.7%)</td>
</tr>
<tr>
<td>Week(^b)</td>
<td>68.3% (57.3%, 80.8%)</td>
<td>83.3% (69.3%, 96.7%)</td>
<td>50.3% (42.0%, 58.0%)</td>
<td>74.5% (74.2%, 74.8%)</td>
<td>84.3% (84.0%, 84.6%)</td>
<td>58.8% (58.4%, 59.2%)</td>
</tr>
</tbody>
</table>

\(^a\)The denominator is the "Twitter full archive" dataset.
\(^b\)The denominator is the union of "Twitter COVID-19 stream" and "Twitter multi-crawler search" datasets.

Figure 3 visualizes the data overlap in different scenarios across the data collection methods.

Figure 3. Data overlapping across data collection methods.

RQ2: How representative of the data collected with each of the 5 data collection methods to the gold standard dataset (i.e. from the Twitter full archive)?

We measured the representativeness of each dataset to the "Twitter full archive" dataset in terms of the Kendall correlations based on ranking of keywords. As shown in Table 7, all the data collection methods in Group 1 have moderate correlations with the "Twitter full archive". The two streaming APIs (i.e. "Twitter sampled stream" and
In this study, we aimed to assess data collection bias among different data collection methods provided by Twitter, identify the representativeness of each data collection method compared with the "Twitter full archive", and test ML models’ sensitivities to data collection bias, through answering three research questions (RQs).

For RQ1, we found that, first, the "Twitter multi-crawler search" can effectively collect more tweets than a single "Twitter search" crawler, even with the same set of keywords, in terms of both data volume and overlapping...
proportions (i.e. covering more samples of the Twitter universe), suggesting that (1) Twitter's internal subsampling strategies might not be consistent across different endpoints due to rate limits, and (2) using multiple crawler and multiple queries is a way to work around the API rate limits. Nevertheless, using multiple crawlers requires researchers to have multiple Twitter accounts, and as Twitter has strengthened its identity verification process, especially for developer accounts (e.g., each phone number can only register a single Twitter account), this presents a challenge. We also found that overlapping proportions between the “Twitter filtered stream” vs. “Twitter search” in Group 1 and the “Twitter COVID-19 stream” vs. “Twitter multi-crawler search” in Group 2 increased when we compared only the tweets that included the top 10 keywords, leading to our RQ2, the representativeness of the tweets collected by different data collection methods.

To answer RQ2, we assessed the Kendall correlations between the "Twitter full archive" benchmark dataset and each of the other five datasets. Among the five other datasets, the “Twitter filtered stream”, "Twitter sampled stream", and "Twitter search" datasets all have moderate correlations, which indicates that even though these crawlers only collect subsamples of the tweets in Twitter data warehouse, these subsamples are still representative and can be used to "identify and track trends, monitor general sentiment, monitor global events, and much more" as claimed in a Twitter API document. Somewhat surprisingly, the "Twitter multi-crawler search" (τ = 0.89) has the same level of correlation to “Twitter full archive” as the "Twitter COVID-19 stream" (τ = 0.89), which shows the effectiveness of the multi-crawlers and multi-queries strategy. The "Twitter COVID-19 stream" was designed based on Twitter's internal COVID-19 Tweet annotation and parameters, which they "believe deliver a comprehensive view of the conversation around this topic." The effectiveness of the "Twitter multi-crawler search" gives us some level of confidence that studies that use the "Twitter multi-crawler search" for other public health studies, where an endpoint like "Twitter COVID-19 stream" does not exist, can still be conducted.

To answer RQ3, we randomly selected eight training samples for building ML models from different parts of the datasets as shown in Figure 2 (B). Among the eight training samples, seven samples were selected either from a single dataset or from overlapping portions within each group, and one sample was selected from the “Twitter full archive” dataset covering the time ranges of both Group 1 and Group 2. We found that (1) in general, models trained on one dataset cannot perform well on the samples from the other group (i.e., time shifts have a significant impact on model performance); (2) CNN models can achieve the highest average performance with relatively reliable consistency when using sample from the “Twitter full archive”. These results suggest that it is necessary to selection training samples from a representative dataset, and as the time progress, it is important to retrain of ML models with new datasets; and (3) traditional ML models such as RF and GBT can only achieve reasonable performance on the samples that they trained with. This indicates that compared with CNN models, RF and GBT are more easily to be overfitted. Thus, when conducting social media analyses, more thorough experimentation and testing of the selected models and their underlying assumptions of the data is necessary.

We also recognized the limitations of our study. First, we cannot recover the “Twitter full archive” dataset back to the time when we collected the other datasets, because of issues such as deleted tweets and suspended accounts. Thus, our measures of overlapping between other datasets to the full archive are only approximates. Second, many other factors that may affect the Twitter data collections such as the number of crawlers running on a single machine (i.e., competing of CPU cycles), and the reliability of the Internet connections among others. Third, many other factors may affect the ML model performance as well such as the sample size of the training samples, data preprocessing methods, hyper-parameter tuning, and data imbalance issues. Fourth, there are many other types of biases in Twitter studies, such as demographic bias and keyword bias. Weeg et al. (2015) mitigated demographic bias of Twitter data by stratifying Twitter users based on geographic distributions. Kim et al. (2016) measured the quality of data collection in two aspects: 1) retrieval precision (i.e., “precision measures how much of the retrieved data is not garbage” and 2) retrieval recall (i.e., “recall measures how much of the relevant data is retrieved”) and proposed a conceptual framework for the filtering and quality evaluation of social data. Those biases and potential methods are worth investigating in future research.

In conclusion, our study assessed the data collection bias, evaluated the representativeness of multiple data collection methods, and tested ML models’ sensitivity to data collection bias. Data and model bias issues are often ignored in social media studies. However, to really use social media such as Twitter as a reliable data source for future research, we must find ways to address (or at least assess) data and model biases.

**Acknowledgment**

This work was supported in part by NSF Award #1734134.
Towards Digestible Digital Health Solutions: 
Application of a Health Literacy Inclusive Development Framework for 
Peripartum Depression Management

Alexandra Zingg, MPH, MS, Tavleen Singh, MS, Sahiti Myneni, PhD

University of Texas Health Science Center at Houston, School of Biomedical Informatics, Houston, Texas, USA.

Abstract
Women of low income and education have lower levels of peripartum depression (PPD) literacy, limiting their ability to recognize symptoms and make informed healthcare decisions. Existing digital solutions and underlying development frameworks for PPD lack an integrative approach addressing health literacy and related disparities. Therefore, we develop an integrative framework for digital content engineering in PPD self-management consisting of (a) user needs analysis, (b) inclusion of eHealth literacy principles (science and health literacy), and (c) mapping user needs to the Behavioral Intervention Technology model. Results revealed that perinatal women seeking mental health care prefer information in multisensory formats, and knowledge needs were identified in areas such as medication management and coping with abnormal results. Results were mapped to eHealth literacy features of whiteboard videos covering essential PPD knowledge, and social media features where patients can articulate information needs. Initial evaluation of proposed features against existing PPD self-management solutions are discussed.

Introduction
Peripartum depression (PPD) affects approximately one out of 10 women in the United States [1], and may cause significant maternal and neonatal morbidity [2]. PPD can affect any woman, but those with characteristics such as belonging to minority groups or of low socioeconomic status (SES) are the most vulnerable [3]. Additionally, women of minority groups such as African-Americans and Latinas are likely to experience disparities in accessing and completing adequate treatment [4]. Digital health technologies have been touted as a viable solution that can improve access to PPD care, and previous research has tested such solutions in the areas of PPD screening [5] and remote delivery of therapy programs [6]. Other areas of PPD care where digital interventions have been explored are self-monitoring of mental health status [7], and providing of social support during pregnancy [8]. Results from these previous interventions show that digital solutions can improve the management of PPD and lead to better mental health outcomes for perinatal women [9].

The indispensable role of health literacy in reducing disparities related to PPD is well-established. For example, low health literacy can adversely impact women’s ability to correctly recognize depression symptoms [10] and to proactively initiate the seeking of help and information for depression and other mental health disorders [11, 12]. Low health literacy is also associated with less understanding of PPD treatment options [10]. And, it has been shown that women with low literacy are likely to experience adverse socioeconomic conditions such as employment insecurity [13], making it more difficult to access mental health care resources. Within the context of PPD, existing work that evaluates digital health technology interventions includes randomized controlled trials [14], feasibility studies [15], and qualitative studies [16, 17] regarding interventions such as applications and text messages. Some of these studies have shown that health literacy is an important factor in women’s engagement with digital information sources. For example, Guendelman and colleagues [17] explored the role of eHealth literacy in disadvantaged perinatal women’s engagement with digital-health-seeking activities, and found that it was positively correlated with higher web-based seeking activities. A qualitative study by Pineros-Leano and colleagues [16] also indicated that clinic staff viewed reduction in literacy barriers as an advantage of having women complete PPD screening evaluations through an electronic tablet format. Clinic staff specifically mentioned the abilities to swiftly adapt screening questions to multiple languages, and to have them read aloud to the patient, as unique literacy components offered by the tablet. Although some of these studies briefly mention examples of how digital tools can address health literacy factors, existing literature underexplores the development and application of content engineering methods guided by health literacy principles for PPD digital health solutions.
While previous research shows that digital interventions can improve PPD health outcomes for vulnerable populations [9], it is now well known that sometimes digital health technologies themselves can create new disparities or exacerbate existing ones [18]. This phenomenon has been termed a “digital divide” in healthcare [19], and it refers to how the benefits of digital health solutions do not always reach all populations, including the ones who can benefit the most from them. In the case of PPD, existing literature [17] indicates health literacy barriers can prevent some women from fully enjoying the benefits of digital health technology solutions. In order to prevent such inequities and facilitate increased comprehension and engagement with digital information resources, the purpose of this study is to outline an integrative content engineering framework to supplement existing design and development processes of PPD digital solutions. To achieve this, we analyze women’s information and technology needs regarding PPD through the lens of health literacy theory. Our aim is to contribute new knowledge towards the production of easily digestible digital content that can increase women’s knowledge regarding PPD, and ultimately result in improved health outcomes for this population. We incorporate principles from the eHealth literacy framework (specifically, learning categories and the literacy domains of health and science) into the behavior intervention technology model. This allows us to propose and outline eHealth literacy features which present essential PPD information in an engaging and minimalist manner (examples: whiteboard videos and social media channels), and which can vastly improve women’s knowledge and understanding of this important condition.

Methods

Theoretical Rationale: eHealth Literacy

Health literacy is a term that has recently undergone a redefinition in August 2020 by the U.S. Department of Health and Human Services’ Healthy People 2030 initiative [20]. The term has been expanded to include both personal health literacy and organizational health literacy. Personal health literacy is “the degree to which individuals have the ability to find, understand, and use information and services to inform health-related decisions and actions for themselves and others” [20], while organizational health literacy is “the degree to which organizations equitably enable individuals to find, understand, and use information and services to inform health-related decisions and actions for themselves and others” [20]. This new definition is meant to highlight the important role of organizations in enabling individuals to reach optimal levels of health literacy.

With the rapid rise in use of digital health technologies, an additional separate term was created to reflect how health literacy fits into this new digital health landscape. The term is electronic health literacy (or eHealth literacy), and is defined as "the ability to search, locate, understand and use health information through electronic resources and use this knowledge to resolve health-related problems" [21]. Some theoretical frameworks have been derived to further elaborate on the concept of eHealth literacy. One of the earliest is Norman and Skinner’s “Lily Model” [21], which establishes six types of literacy (traditional, information, health, science, media, and computer) as its subcomponents. Chan and Kaufman [22] expanded on this framework and created the “eHealth Literacy Framework” by addition of cognitive elements from “Bloom’s Taxonomy” [23], which is extensively used to plan and execute educational objectives. It is a collection of action words that describe how a user processes and synthesizes new knowledge [23] and contains six categories of learning, from the simple level of remembering to the complex level of creating. Additional literacy skills and factors in predicting patient’s use of health information technology were also added to include privacy and security concerns, need for information and self-management, computer anxiety, and convenience and advantages of information from health professionals [24,25].

While there are multiple health literacy frameworks, Chan and Kaufman’s offers the most user-ready, systematic process of attributing different literacy domains and cognitive tasks to digital health features. It allows developers to determine at what cognitive level each health literacy domain is applied in digital features. In this paper, we utilize this framework in conjunction with our Digilego framework, which comprises of user needs analysis and BIT model [26] to develop theory-driven digestible digital solutions for PPD management.

Figure 1 presents an overview of our integrative framework. The main components of the framework - (a) mixed-methods needs analysis, (b) theory mapping, and (c) content specifications are as discussed below.
Figure 1- Incorporation of eHealth Literacy Framework into PPD-specific Digital Health Technologies

User Needs Analysis: We conducted a mixed-methods user needs analysis consisting of focus groups, interviews, and social media analysis. While these datasets were originally produced during previous work [27, 28], in this study we use them to inform the mapping of health literacy digital features to the condition of PPD. Two focus groups were conducted with a total of nine patients seeking prenatal care at a Maternal and Fetal Medicine clinic in Houston, TX. Each session lasted approximately thirty minutes. Additionally, ten patients were individually interviewed. The purpose of the focus groups and interviews was to discover themes regarding the use of digital technologies in the self-management of mental health during the peripartum period. They followed a semi-structured format in which question guides were used to facilitate open discussion. Question guides included the topic of educational content and design, the information they receive from currently available products, and their current knowledge of depression (example questions: “What kind of information do you actually get from these apps?”; “How would you like any educational materials of the app to be prepared?”). Ground theory analysis was used to extract themes from focus group and interview recordings, in which opening coding was done on a line-by-line basis to derive the concepts being brought up by the participant. This was followed by axial coding, in which patterns and relationships between concepts were delineated to produce major themes. Information from our focus groups and interviews was complemented with analysis from social media data. We chose to analyze online PPD forums from the websites of BabyCenter [29] and What To Expect [30]. Both of these websites are highly popular with pregnant women, and host very active social forums. While women do not have to be diagnosed with PPD in order to participate in the PPD market, a total of 5,532 posts were extracted from What to Expect, and 56,483 posts from BabyCenter. We manually labeled 850 user posts with the following five categories: Breastfeeding, Family and Friends, Medications, Symptom Disclosure, and Social Support. Similar to our focus groups and interviews, these categories were derived using grounded theory analysis. We then applied machine learning (ML) models to scale up the qualitative labels to the entire dataset. We specifically applied three ML classifiers to our labeled dataset: Logistic Regression (LR), Random Forest (RF), and Support Vector Machine (SVM). Our automatic text classification was multilabel, meaning multilabel categories were allowed per each user post. To achieve this, we applied the one-vs-the-rest multilabel strategy, whereby binary classifier is trained per category, and then categories are fitted against each other. Our automatic text classification was completed using Python’s sciKit-learn package [31]. The classifiers were evaluated with the measures of recall, precision, and F1 score. After evaluation, we applied the best performing model to the previously unlabeled dataset from BabyCenter.

Mapping Health Literacy to PPD: We then employed Mohr’s Behavior Intervention Technology (BIT) Model to incorporate health literacy principles into PPD digital features. This model asks the five questions of Why? How? What? and When? This model can be easily applied to not only the general technology development process but also to specific components of the technology [32]. To help us answer the question of How?, we have used the literacy skills and learning categories previously described as part of the eHealth Literacy Framework. We then finalized a set of PPD literacy features to be integrated with our digital solution, MomMind [28]. An example that illustrates this mapping process is: the task of articulating a question about perinatal mental health through the literacy feature of social media forums is related to the eHealth Literacy Framework cognitive task of analyzing information within the literacy domain of health.

Initial Evaluation: We have also conducted an initial market evaluation of our proposed features against literacy features in existing PPD applications, as measured by their reading grade level and cognitive task complexity. To
evaluate the reading grade level of the application’s content, we used the Flesch-Kincaid readability test [33]. This is a measure that is widely used, and it is based on a mathematical equation factoring the text’s total words, syllables, and sentences. Scores of the test range from 0-100, with lower scores indicating a more difficult reading level and higher scores indicating an easier reading level. Cognitive task complexity was determined through the six categories of learning described in the eHealth literacy framework (i.e., understanding, applying, evaluating). This evaluation can help guide future researchers and developers in assigning the appropriate complexity level of their health literacy content.

**Results**

*a) User Needs Analysis*

**Focus Groups and Interviews**

The majority of our focus group and interview participants were young: 10 were 25-34 years old, five were 18-24 years old, and four in the category of 35-44 years old. The ethnical composition of our sample was: seven identified as Hispanic, six as Black, four as White, one as Asian, and one as Other. One participant had an education level of some high school, three participants were high school graduates, five participants had some college credit, five had an associate’s degree, four had received technical training, and one had a bachelor’s degree. 14 out of 19 participants were low-income women. The main source of pregnancy information for participants was the internet (n=17). 11 participants reported seeking additional information from non-digital sources like their doctors or friends and family. Our focus groups and interviews revealed various literacy themes, including preferred formats for educational materials. The majority of participants preferred to consume educational content through dynamic formats (e.g., slide shows, videos, audio) that would make the process of understanding the content more amenable to them. This is illustrated in Comments 1 and 2. Such formats were also preferred because they are a good fit for the busy schedules of expectant and new mothers.

*Comment 1*: “… kind of like how it is now through a slideshow or, you know, just kind of something similar to that. Short videos.”

*Comment 2*: “If you could actually listen to them instead of reading, it’s better because I’m a mother so I don’t always have time to read.”

Participants also mentioned the importance of separating topics into sub-topics for a well-paced presentation of the material (Comment 3). This is in line with previous research that suggests separating information into blocks of about seven units makes it easier for the learner to retain information in memory [34].

*Comment 3*: “Well, when I was in school, I started off like the main topic and then I have bulletins that will help break down. You know, okay, like if it was for a symptom it brings it down to like, what I should do, or how I should treat it, any medication... broken down to the bullets or sections.”

Participants also expressed the at times frustrating experience of following up with their doctors after having read information from digital sources (e.g., social media). It was difficult for them to discern when to trust these sources, and when to contact their doctors if such sources indicated a possibly abnormal event in their pregnancy. This indicates a barrier for participants in using the information they receive from non-traditional sources, as Comment 4 shows.

*Comment 4*: “Sometimes I’ll come to my doctor (with the app) and show ‘Look, this says this is not normal’ and they’ll say ‘Oh no, don’t trust those things’. Well, then I don’t want to bug you every five minutes for the nine months we get to go through this. I was told this would not happen to me.”

**Social Media Analysis**

Evaluation of our selected ML models indicated that the RF model had the best performance (Table 1). Therefore, we applied this model to predict categories for user comments in our unlabeled dataset of 56,484 user posts from BabyCenter. The category used most by our chosen model was “Social Support” with 50,337 comments, followed by “Medications” with 10,499. “Symptom Disclosure” was applied to 6522 posts, and “Family and Friends” to 2275 posts. The least used category was “Breastfeeding” with 532 comments. The high frequency of the “Social Support” category indicates that women use channels such as online social forums both to receive and provide support regarding PPD at various levels: emotional (words of encouragement), informational (sharing of scientific and health concepts regarding PPD), and instrumental (practical tools to help manage PPD, such as relaxation techniques). Therefore, it is
important that discussion participants are able to critically process the information from such channels by having sufficient baseline knowledge about the condition. A specific health literacy component that we can derive from this result is that all women need a simple introduction to PPD that covers essential information such as its definition, symptoms, and possible treatments. The high number of “Medications” post also indicates that women who seek PPD care need educational materials that will help them understand practical topics such as how to better manage their medication (i.e., different medication options, usual dose, frequency) and the medication effects on breastfeeding and baby’s health.

### Table 1 - Automatic Text Classification Model Performance

<table>
<thead>
<tr>
<th>Category</th>
<th>Logistic Regression</th>
<th>Random Forest</th>
<th>Support Vector Machine</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
<td>F1 Score</td>
</tr>
<tr>
<td>Family and Friends</td>
<td>0.89</td>
<td>0.59</td>
<td>0.71</td>
</tr>
<tr>
<td>Medications</td>
<td>0.95</td>
<td>0.73</td>
<td>0.83</td>
</tr>
<tr>
<td>Symptom Disclosure</td>
<td>0.71</td>
<td>0.68</td>
<td>0.69</td>
</tr>
<tr>
<td>Social Support</td>
<td>0.62</td>
<td>0.85</td>
<td>0.72</td>
</tr>
<tr>
<td>Breastfeeding</td>
<td>1.00</td>
<td>0.26</td>
<td>0.41</td>
</tr>
</tbody>
</table>

**b) Theory mapping**

Table 2 shows our mapping of the BIT model to health literacy components for PPD. The goal of incorporating health literacy components into the content of our educational platform is to make our content as learnable as possible to the majority of women. Ultimately, this will allow them to use new knowledge to make the best decisions about their mental health (i.e., deciding to disclose symptoms to their doctors) and sustain positive health behaviors (i.e., routine mental health screening). To accomplish the goal of making our content highly learnable, we will follow results from previous research on education and psychology. For example, studies have shown that presenting content to the user in short segments helps improve long-term memory retention [34, 35]. Therefore, our content will be presented in brief whiteboard videos. As also indicated by our participant demographics, the education level of our participants ranged from some high school to a college degree. To assure that our materials are written at a literacy level that can be well understood by all, we have selected to write our educational modules at a reading level of eighth grade.

### Table 2 - Behavioral Intervention Technology Model Mapping to Health Literacy for PPD

<table>
<thead>
<tr>
<th>BIT Component</th>
<th>Mapping to PPD (Examples)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Why</td>
<td>Improving PPD health literacy by increasing user knowledge. Aim is for the user to access and retain educational materials about PPD.</td>
</tr>
<tr>
<td>How (conceptual)</td>
<td>The eHealth Literacy Framework, which combines six categories of literacy (health literacy being one of them) and six categories of learning. Overall, this framework represents 36 unique ways to engage users into improving their health literacy. Example: Participant’s health literacy on PPD can be improved by helping them understand the differences between various treatment options, allowing them to make the best decision with their doctors.</td>
</tr>
<tr>
<td>What</td>
<td>Educational Modules: Separation of content into brief multiple units Allow user to consume content at their own pace (ability to pause, stop) Providing user opportunities to articulate what they have learned (interactivity with providers and peers, journaling)</td>
</tr>
<tr>
<td>How (Technical)</td>
<td>Multimedia storytelling techniques (Whiteboard videos, presentations) Maximum of eighth grade reading level (as determined by the Flesch-Kincaid reading test) Avoidance of technical jargon Aesthetically pleasing fonts and colors Goal Setting Progress Monitoring</td>
</tr>
<tr>
<td>When</td>
<td>Educational modules delivered once a week to prevent user fatigue while maintaining user engagement.</td>
</tr>
</tbody>
</table>

**c) Content specifications**
Table 3 shows a granular mapping of the eHealth literacy framework to proposed PPD digital modules in our MomMind digital platform. In this mapping, we have concentrated on the specific literacy domains of health literacy and science literacy. Our results indicate that our platform has the potential to improve our users’ literacy levels in these domains by helping them achieve the cognitive tasks of understanding important PPD health topics and science concepts.

**Table 3- Mapping of the eHealth Literacy Framework to PPD Digital Modules**

<table>
<thead>
<tr>
<th>Digital Features (MomMind Modules)</th>
<th>Applicable Literacy Domain(s)</th>
<th>Literacy Techniques</th>
<th>Cognitive Task(s)</th>
<th>Media Options and Example User Interactions</th>
</tr>
</thead>
<tbody>
<tr>
<td>“My Diary”: journaling feature where user can freely document their thoughts and emotions.</td>
<td>Science, Health</td>
<td>Writing diary entry</td>
<td>Remembering events; Analyzing events</td>
<td>Text, Audio, Pictures</td>
</tr>
<tr>
<td>Example: User writes diary entry about activity they did with baby and includes picture</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>“Mom Talk”: Social forum component where participants can discuss PPD topics</td>
<td>Science, Health</td>
<td>Reading other users’ post; Writing by creating posts or replying to posts</td>
<td>Analyzing other’s posts; Evaluating information in other’s post</td>
<td>Text, Audio, Video</td>
</tr>
<tr>
<td>Example: User discusses an article about healthy sleeping habits with forum peers</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>“My Care”: represents the clinical profile of the user, including provider profile and medication information</td>
<td>Health</td>
<td>Reading provider’s answers; Writing questions to providers</td>
<td>Analyzing answers from providers</td>
<td>Text</td>
</tr>
<tr>
<td>Example: User can learn about PPD medication management by asking their doctor about potential side effects through the “Ask your provider a question!” form.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>“My Library”: repository of evidence-based educational PPD information</td>
<td>Science, Health</td>
<td>Viewing educational content</td>
<td>Understanding PPD concepts presented in multimedia educational videos</td>
<td>Text, Video, Audio</td>
</tr>
<tr>
<td>Example: User can learn about hormonal changes in the peripartum period through viewing a whiteboard video.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>“How am doing today?”: PPD survey repository</td>
<td>Health</td>
<td>Reading survey questions; Selecting best answer</td>
<td>Understanding survey questions; Evaluating survey results</td>
<td>Text</td>
</tr>
<tr>
<td>Example: User completes PPD screening presented in written format</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

d) *Initial Evaluation*

Our initial market evaluation of literacy features in our proposed digital platform for PPD management against five selected commercial applications for PPD management [36-40] showed that most applications addressed health literacy through educational modules and brief tips for PPD symptom management (Table 4). It was found that two of the applications had content written for the college reading level, one had content written for the high school reading level, and one had content written for the eighth-grade level. The content of MomMind was written at the sixth-grade level, making it the easiest to read. These results indicate that most available market applications for PPD management have content which will be most useful to women of high education levels. The application which had the highest level of cognitive task (applying) was PPD Act, which presented weekly practical tips to its users. Three applications (Postpartum Depression, Postpregnancy Recovery, and MomMind) had the lowest complexity level of cognitive task from the eHealth literacy framework by presenting PPD information that the user can recall from their long-term memory if needed.
<table>
<thead>
<tr>
<th>Application Name</th>
<th>Example of Digital Feature related to PPD Health Literacy</th>
<th>Sample Health Literacy Content</th>
<th>Flesch-Kincaid Reading Grade Level</th>
<th>Cognitive Task</th>
</tr>
</thead>
<tbody>
<tr>
<td>Moment Health</td>
<td>“Community”: Link to Facebook group where participants can write posts to discuss PPD topics and ask questions to fellow new parents.</td>
<td>Not Available</td>
<td>Not Applicable</td>
<td>Understanding PPD-related information from user posts</td>
</tr>
<tr>
<td>PPD Act</td>
<td>Weekly tips to better manage PPD symptoms.</td>
<td>“Protected sleep time is critical for women that suffer from PPD. Aim for at least 5 hours of uninterrupted sleep at night”</td>
<td>10.1</td>
<td>Applying weekly tips into routine behavior</td>
</tr>
<tr>
<td>MGHPDS</td>
<td>Link to the Massachusetts General Hospital’s Center for Women’s Mental Health website, containing recent PPD research articles.</td>
<td>“Although new mothers describe breastfeeding as a meaningful and fulfilling aspect of caring for their infants, breastfeeding is also a common source of stress and anxiety.”</td>
<td>13.1</td>
<td>Analyzing information from research articles</td>
</tr>
<tr>
<td>Postpartum Depression</td>
<td>PPD education modules which present information in a Question-and-Answer format, and modules with recommendations to alleviate PPD symptoms.</td>
<td>“Postpartum depression is a mood disorder that can affect women after childbirth. Mothers with postpartum depression experience feelings of extreme sadness, anxiety, and exhaustion that may make it difficult for them to complete daily care activities for themselves or for others”</td>
<td>13.7</td>
<td>Remembering information from PPD education modules</td>
</tr>
<tr>
<td>Post Pregnancy Recovery</td>
<td>Topic modules concentrating on physical recovery in the postpartum period; one module is dedicated to PPD information.</td>
<td>“The postpartum period refers to the first six weeks after childbirth. This is a joyous time, but it’s also a period of adjustment and healing for mothers. During these weeks, you’ll bond with your baby and you’ll have a post-delivery checkup with your doctor”</td>
<td>8</td>
<td>Remembering information from PPD topic modules</td>
</tr>
<tr>
<td>MomMind</td>
<td>“My library” (described in Table 2).</td>
<td>“In fact, 6% of pregnant women and 10% of new mothers experience anxiety. This can occur along with depression, or on its own.”</td>
<td>6.8</td>
<td>Recall information from library content</td>
</tr>
</tbody>
</table>

**Discussion**

Our user needs analysis provided us with an in-depth dwelling into the world of perinatal women who are seeking mental health resources and the hurdles they face in their journey. Our target population (low-income women seeking mental health care) has an educational level of mostly some college, but the lowest education level achieved was some high school. Most were also low-income and of minority groups. Overall, these characteristics indicate to us that any educational materials presented to this population should be at a maximum reading level of 8th grade, to assure that the materials can be understood by everyone. We also found that this population was most interested in learning about topics that served as gateways into further discussion and information about mental health and PPD (breastfeeding, family environment, medications). Our user needs analysis could have been strengthened by including in our focus group and interview samples women of different backgrounds who have been diagnosed with PPD.

Our theory mapping of the BIT model and the eHealth literacy framework to literacy features for PPD allowed us to model the relationship between our target users, our content, and the digital environment. This mapping suggests that
cognitive tasks such as remembering, analyzing, and applying PPD information would increase our user’s PPD health and science literacy. From this mapping, we derived the following two categories of health literacy features that should be included in digital solutions for PPD self-management:

a) Brief, multimedia educational materials: As indicated by our participant’s focus group and interview comments, participants prefer to receive information in a manner that is not just text. They specifically mentioned video and audio as formats that are engaging. They also suggested information should be broken down into bite-size chunks that can be consumed in short spans of time. Another conclusion we can make from our user needs analysis is that many women lack essential knowledge on PPD, such as its definition, symptoms, and available treatment. Therefore, our proposed application will contain brief whiteboard videos that will present women with this essential knowledge on PPD.

b) Channels where the user can articulate and analyze information: The combined results of our social media analysis, focus groups, and interviews revealed that the ability to discuss PPD information with peers and providers was an important component in increasing women’s health literacy. For example, our social media analysis revealed that online PPD forums was a very important source of information where women would share knowledge on a wide range of PPD areas, from recommendations on clinical treatment to sharing of local support resources. For this reason, in our proposed digital platform we include the features of a social forum where women can interact with fellow patients from the clinic they are attending, as well as the ability to have bidirectional communication with their providers to ask questions about the peripartum period and PPD.

The strength of our mapping process is that it is based on well-established literacy and technology theories. However, a main limitation of our study is lack of extensive evaluation of our selected literacy features with PPD stakeholders, including patients and clinicians. Additionally, our literacy features should be evaluated across users with high and low levels of health literacy. As part of future steps in our research program, we will focus on implementing these features and evaluating them in formative and summative evaluation studies to understand the effects of our framework on patient engagement, psychosocial outcomes, and PPD self-management capabilities. Our study adds important knowledge that is currently missing in existing literature: it explores the theoretical application of health literacy principles in the development and application of content engineering methods for digital health solutions in PPD management.

Conclusion

Health literacy is an important factor in empowering patients to make the best health decisions and reaching optimal health outcomes. With the rapid rise in use of electronic sources to distribute health information, it is critical that developers and providers incorporate electronic health literacy factors into digital health solutions to assure that such sources can be easily understood and used by the intended target audience. Such literacy factors can range from media type through which the information is delivered to the reading level at which the information is written. The integrative framework we describe in this study allows us to guide the work of future researchers, digital health designers, and developers to determine the ways in which literacy principles and digital features should be engineered. Ultimately, this framework can result in digital health technologies that promote equity in access to health resources.

Acknowledgements

We would like to thank patients and staff at the UTPhysician’s Women’s Center at TMC who participated in this study for their time and insights. Research reported in this publication was partly supported by the National Library of Medicine of the National Institutes of Health under award numbers 1R01LM012974-01A1. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.
References


34. Miller GA. The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychological Review. 1956, 63(2), 81–97. https://doi.org/10.1037/h0043158
Racial Representation Analysis in Dermatology Academic Materials

Girmaw Abebe Tadesse, PhD, Celia Cintas, PhD, Roxana Daneshjou, MD, PhD, Kush R. Varshney, PhD, Peter Staar, PhD, Skyler Speakman, PhD, Kenya Andrews, Chinyere Agunwa, PhD, Justin Jia, Elizabeth Bailey, MD, MPH, Ademide Adelekan, MD, Jules B. Lipoff, MD, Ginikanwa Onyekaba, MD, Veronica Rottenberg, MD, James Zou, PhD

1IBM Research – Africa 2Stanford University 3IBM Research – T. J. Watson 4IBM Research – Europe 5University of Pennsylvania 6Memorial Sloan-Kettering Cancer Center

Introduction

Because skin disease appears visually different across skin tones, educational materials depicting diverse skin tones are required for a well-trained healthcare workforce1–5. However, dermatology textbooks, lecture notes, and published literature lack adequate diversity in skin color representations. Massie et al. analyzed skin tones in the New England Journal of Medicine from 1992 to 2017 and found that only 18% of the images represented non-white skin.3 Adelekun et al. assessed skin tones in the top general dermatology textbooks and found darker skin tones significantly under-represented.5 The lack of representation in educational materials may translate to the clinical realm, where skin cancer diagnoses (e.g., melanoma, squamous cell carcinoma) are significantly delayed in patients of color, leading to increased morbidity and mortality6. The COVID-19 pandemic has further highlighted inequities4; manually annotated published photos of COVID-19 cutaneous findings and found images depicting darker skin lacking. Unfortunately, manual skin tone annotation is not feasible for a large corpus of dermatology education materials due to its labor-intensive nature, operator visual fatigue, and intra-/inter-observer error related to category assignment for different skin tones1,4. Automatic assessment of skin tone representation would significantly aid in identifying bias in educational materials. Previous work7 used Individual Typology Angle (ITA) to approximate skin tones in curated image datasets used as benchmarks and found they under-represented darker skin tones. We extend this work to process images in the wild from off-the-shelf textbooks and output representations of skin tone, which allows domain experts (e.g., dermatologists) to visualize biases in the data.

Method

The overview of the method is shown in Fig. 1. The main components are automatic ingestion of traditional academic materials (e.g., textbooks in .pdf), parsing of different entities (e.g., figures), extraction of skin images, and estimation of skin tones.

Figure 1: Overview of the proposed framework. Academic materials (e.g., in pdf format) are fed as input to our pipeline, which are ingested using a corpus conversion service8 resulting document entities (e.g., images and tables) parsed into a structured format (e.g., JavaScript Object Notation - JSON). Images are then cropped out using the annotations, among which skin images are selected. Segmentation of skin pixels is done followed by estimation of skin tones using a pre-trained ResNet-18 framework with the last layer fine-tuned to output “light” vs “dark” skin tones.

Data and Results:

We evaluated our framework using images extracted from four dermatology textbooks: Rook’s textbook of dermatology9, Bologna 4e10, Fitzpatrick Color Atlas 8e11 and Fitzpatrick Dermatology in General Med 9e12 resulting 28159 images collected from 416 chapters. The detection of skin images is performed with a leave-one-textbook-out validation and results in AUC of 0.98 ± 0.02 and F1 score of 0.94 ± 0.03 using XGBoost classifier. After
skin image segmentation is employed using pixel-level intensity values and region information, we aim to identify the skin tone, i.e., light vs. dark. To this end, we employed both handcrafted (e.g., HOG, ITA) features and raw pixels to train the models. Both traditional tree-based classifiers (e.g., gradient boosting) and pre-trained ResNet are experimented with five-fold stratified cross validation. The best performance achieved by handcrafted features is using a Gradient boosting classifier with $0.85 \pm 0.03$ AUC and $0.89 \pm 0.01$ $F_1$ score. A ResNet framework (pre-trained with natural images and finetuned to output "light" vs "dark" skin tones) outperformed the baseline classifiers using selected pixels after masking, i.e., $0.89 \pm 0.01$ AUC and $0.95 \pm 0.00$ $F_1$ score.

Discussion: Disparities in dermatological outcomes may be related to inequities in dermatological education, particularly the lack of darker skin images in educational materials used to train dermatologists and primary care physicians. Thus far, efforts to quantify biases in these materials have been done by hand, which is labor-intensive and impractical for large-scale application. Thus, we aim to develop a standalone online tool that enables a domain-expert to upload a given academic material and obtain distribution of images across different skin tones. This provides first-hand awareness of potential bias existing in these materials and thereby a chance to improve the imbalance. Since all images from a dermatology textbooks or educational materials are not skin related, the first step aims to discard non-skin images. Skin tone estimation needs to also avoid lesions areas of the skin image (e.g., lesion pixels in light skin image might resemble dark skin tone) and unnecessary background and foreground components (e.g., clothes). Results show encouraging performance in detecting skin images and classifying skin tones. The findings confirmed the reports of lack of dark skin images from manual studies of dermatology materials as only an average of $10.15\%$ of dark skin images (compared to light skin images) is found in the four validated textbooks. Future work aims to pilot the developed tool in different geographical locations. We envision this technology as a tool for dermatology educators, publishers and practitioners to quickly assess their educational materials, which could be scaled to other domains (e.g., history) to automatically identify lack of representation in, e.g., race and ethnicity.

References

Feasibility of a Self-Administered Elder Abuse Digital Screening Tool for Use in Emergency Department Settings

Fuad Abujarad, PhD, MSc1, Esther Choo, MD2, Michael V. Pantalon, PhD1, Karen Jubanyik, MD1, James Dziura, MPH, PhD1, Chelsea Edwards, BSc1, Gail D'Onofrio, MD, MS1, Thomas M. Gill, MD1

1Yale School of Medicine, New Haven, CT, USA; 2Oregon Health & Science University, Portland, OR, USA

Introduction

Elder abuse is a complex, national problem where as few as 1 in 24 cases may become known to programs and agencies1. Existing methods for identifying victims are limited due to underreporting and varying rates of known reporting and there is a need for a feasible and acceptable solution to facilitate and improve identification of elder abuse cases2. We developed a web-based digital health tool delivered on tablets to assist older adults in self-identifying and self-reporting elder abuse3. The VOICES tool was created to be self-administered by the patients themselves and offers a private, personalized experience based on answers provided by the patient. The tool includes interactive multimedia elements such as animations, videos, and automated text-to-speech (Figure 1). VOICES aims to encourage self-disclosure of abuse by guiding the user through a screening process followed by educational and motivational modules. The educational module details common types of abuse, services available and what may happen after disclosure. The motivational module encourages the user to recognize potential harmful behaviors and conveys that the consequences of abuse could be avoided by reporting abuse to a professional. In this abstract, we will discuss results from our ongoing study to evaluate the feasibility of the digital health tool VOICES with older adults age 60 or older in the emergency department.

Methods

Design: We conducted a feasibility study of the digital health tool VOICES, as a self-administered elder abuse screening tool at a busy emergency department.

Settings: Participants were recruited from the St. Raphael Emergency Department in New Haven, CT. All procedures were approved by the Yale University Human Investigation Committee (IRB).

Participants: We recruited cognitively intact and community-dwelling individuals 60 years or older who read and spoke English. We sought diversity in age, race, and sex to match the diverse patient population that normally visit the study site. Participants were compensated $20 for their time.

Intervention: The entire study session was conducted privately for patient confidentiality. A research assistant (RA) was available and ready to answer any questions if necessary. The session (one visit) took on average 45 minutes to complete including: consent, demographics questioner, tool use on iPad, and post-survey. After obtaining consent the RA helped the older adult with putting the headphones; oriented them to the iPad and explained the process that followed. First, participants completed a demographic survey. Next, they engaged in the VOICES tool on the iPad. Following completion of the VOICES tool, participants were asked to answer a series of questions validated by our research team on a 5-point Likert scale gauging their feelings regarding feasibility, satisfaction, ease of use and appropriateness of the tool (Figure 2).

Measurement: The study data came from the following sources: (1) participant demographics data, (2) participant responses given to questions during VOICES (abuse screener, self-identification, self-reporting), (3) participant responses on the post-use-survey, and (4) observations made by the RA on enrollment, and VOICES use.

Outcomes: The study was focused on three important outcomes of feasibility including: acceptability, implementation, and practicality.
Results
The participants included 500 older adults ranging in age from 60 to 98.8 years with a mean age of 72.8 years. The majority of participants were female, white, not employed, with high school education or higher, and 65.4% used the internet at least 1-2 times per week. Out of all participants, 98.8% of patients used VOICES to completion. Participants completed the VOICES tool with an average of 8.25 minutes (SD=3.57) and 90.4% participants were satisfied (mean=4.5, SD=0.77) with the time needed to complete the VOICES tool, on a 5-point Likert scale ranging from 1 being “Very Unsatisfied” and 5 being “Very Satisfied”, and 94.4% of participants agreed that the tool was easy to use (mean=4.6, SD=0.64) on a 5-point Likert scale (1: “Strongly disagree”; 5: “Strongly agree”). When asked about the perceived appropriateness of using VOICES to learn about elder abuse, 94.4% agreed VOICES was appropriate to use (mean=4.6, SD=0.64). Also, 93.6% of participants (mean=4.6, SD=0.68) agreed digital screening would maintain privacy and anonymity, and 79.2% felt safer after using VOICES (mean=4.3, SD=0.85).

Conclusion
The results of the feasibility study show that older adults found the VOICES digital health tool to be acceptable, practical, and satisfying. Patient satisfaction was very high, a clear indication of the tool’s acceptability. We found that the tool can be highly implementable since we had a high rate of execution success when measured by the number of participants who used the tool to completion. Participants felt comfortable with the privacy given by the tool and were satisfied with how long it took to complete. Overall, we found that using VOICES to screen for elder abuse in the emergency department setting is practical and feasible. With an implementable, feasible tool to facilitate screening we aim to overcome current barriers with identification. We believe that the VOICES tool is best adapted into an existing emergency department workflow to complement existing protocols for identifying elder abuse, rather than acting as a replacement to current protocols. Further research is needed to evaluate the efficacy of the tool and to explore the use of VOICES in other health care settings outside of the emergency department, such as in primary care settings. Future work involving this tool will include cognitively impaired individuals as well as Spanish-speaking individuals to evaluate previously excluded populations from this study.

This work was supported by the HHS Office of the National Institute on Aging under R01AG060084.

References
Exploring Automatic Summarization of the Hospital Course

Griffin Adams, MS¹, Emily Alsentzer, MS², Mert Ketenci, MS¹, Jason Zucker, MD¹,
Noémie Elhadad, PhD¹
¹Columbia University, New York, NY, US;
²Harvard-MIT Health Sciences and Technology, Cambridge, MA, US

Introduction

Automatic patient record summarization has been proposed to support clinicians in multiple scenarios, from making sense of a patient’s longitudinal record over long periods of time and multiple interactions with the healthcare system, to synthesizing a specific visit’s documentation. Here, we focus on hospital-course summarization, synthesizing a patient’s specific inpatient visit, from admission to discharge. A well-documented hospital course summary is crucial for continuity of care and patient safety post discharge. To motivate future research in multiple, self-contained directions, we distill task-specific characteristics to a few salient, standalone takeaways. For each, we provide evidence in the data and/or literature, before proposing implications of these findings on model development and evaluation.

Methods

Given the clinical documentation available for a patient hospitalization, our task of interest is to generate a text that synthesizes the hospital course in a faithful and concise fashion. The hospital-course summary tells the story of the patient’s admission; it conveys what was done to the patient during the hospital admission and why¹,². To carry out analyses, we create a large-scale summarization dataset, CLINSUM, using the notes from all hospitalizations between 2010 and 2014 at Columbia University Irving Medical Center (CUIMC). We extract the “Brief Hospital Course” (BHC) section of the discharge summary as the proxy (gold-standard) summary, while treating the preceding “Admission”, “Progress”, “Consult” notes as the source documentation. Table 1 shows corpus statistics.

To characterize the task and dataset, we develop several extractive baselines, and report performance in Table 2 on a held-out validation set. Oracle models have access to the ground-truth reference and represent a nice upper bound for extraction. Here, we define different sentence selection criteria for our oracle methods. ORACLE TOP-K: Take sentences with highest $R_{12}*$ vis-a-vis the reference until a target token count is reached; ORACLE GAIN: Greedily take source sentence with highest relative $R_{12}$ gain conditioned on existing summary. Extract sentences until the change in $R_{12}$ is no longer positive; ORACLE SENT-ALIGN: For each sentence in reference, take source sentence with highest $R_{12}$ score; ORACLE RETRIEVAL: For each sentence in reference, take reference sentence from train set with largest BM25 score; and ORACLE SENT-ALIGN + RETRIEVAL: For each sentence in reference, take sentence with highest $R_{12}$ between ORACLE RETRIEVAL and ORACLE SENT-ALIGN. We also provide three extractive baselines: RANDOM - extracts random sentences until summary reaches target word count (average summary length); LEXRANK - selects the top-k sentences with largest LexRank score until target word count is reached; and CLINNEUSUM - a variant of the Neusum model adapted to the clinical genre³. CLINNEUSUM is a hierarchical LSTM-based network trained on ground-truth labels derived from ORACLE GAIN.

Results

Below, we distill dataset characteristics into five standalone groupings. (1) Summaries are mostly abstractive with a few long segments of copy-pasted text. CLINSUM summaries appear extractive according to a coverage (0.83 avg / 0.13 std) and a very density (13.1 avg / 38.0 std)⁴. Yet, there is large variance within summaries. This directly affects the performance of a supervised extractive model, whose selection capability degrades as summary content transitions from copy-paste to abstractive. (2) Summaries are concise yet comprehensive. Hospital course summaries are dense with medical entities regarding disorders, labs, procedures, and medications. These entities are well-distributed across the source notes and relations are often not explicit. Collectively, this makes it a difficult multi-document task, which calls for a knowledge-intensive, domain-specific approach to assessing faithfulness and factuality. (3) Summaries have different style and content organization than source notes. Hospital course summarization involves not only

¹The average of ROUGE-1 & ROUGE-2 F-1 scores.
massive compression, but a large style and organization transfer. Source notes are written chronologically yet the way clinicians digest the information, and write the discharge summary, is largely problem-oriented. (4) Summaries transition abruptly from topic to topic without much cohesion. Clinical text, particular when arranged around a patient’s problems (problem-oriented), exhibits frequent, abrupt topic shifts, as demonstrated by very sparse entity grids (i.e., very few lexical chains). This leads to lower next-sentence prediction scores (NSP), as compared to summarization datasets in other domains (e.g., news articles). We also show that ROUGE is not correlated with local coherence. (5) The BHC sections are silver standard summaries. Discharge summaries and their associated BHC sections are frequently missing critical information or contain excessive or erroneous content.

Discussion and Conclusion

In this abstract, we construct a large-scale multi-document dataset of proxy summaries from the clinician-authored “Brief Hospital Course” section of the discharge note. Based on a comprehensive analysis of the dataset and the literature, we identify a number of implications for this complex task on future research: (1) BHC summarization models must handle abrupt transitions between extractive and abstractive text; (2) the presence of densely packed entities calls for a knowledge-intensive, domain-specific approach to ensuring and assessing the faithfulness and factuality of generated summaries; (3) our oracle analysis suggests that “retrieve-edit” frameworks are well-suited for the task of hospital course generation; (4) as opposed to coherence models based on lexical chains, we argue for the development of problem-oriented models of coherence, which are associative in nature, and reflect a deeper knowledge about the relationship between disorders, medications, and procedures; and (5) noisy reference summaries can harm model performance, but this could be mitigated by dataset filtering or reference-free evaluation approaches.

References

Multi-site Evaluation of Longitudinal Changes in Ejection Fraction in Heart Failure Patients Through Data-driven Phenotyping

Prakash Adekkanattu, PhD1, Jennifer A. Pacheco, MS2, Joseph Kabariti, MS1, Daniel J. Stone, BS3, Yue Yu, Ph.D3, Parag Goyal, MD MSc1, Faraz S. Ahmad, MD MS2, Guoqian Jiang, MD PhD3, Yuan Luo, PhD2, Luke V. Rasmussen, MS2, Pascal S. Brandt, MS4, Zhenxing Xu, PhD1, Jie Xu, PhD1, Fei Wang, PhD1, Natalie C. Benda, PhD1, Thomas R. Campion, Jr, PhD1, Jyotishman Pathak, PhD1

1Weill Cornell Medicine, New York City, NY, 2Northwestern University, Chicago, IL, 3Mayo Clinic, Rochester, MN, 4University of Washington, Seattle, WA

Abstract

In a multi-site retrospective study using electronic health records, we investigated longitudinal changes in ejection fraction (EF) values in patients diagnosed with heart failure (HF) and observed significant variations in comparison with a previous study based on HF registry data.

Introduction

Heart failure (HF) is a major public health problem in the US and around the world. Left ventricular ejection fraction (EF) is a critical measurement used in the diagnosis, prognosis, and treatment of patients with HF. Patients with HF have been categorized into three sub-phenotypes based on their EF: heart failure with reduced ejection fraction (HFrEF) with an EF value less than 40%; heart failure with preserved ejection fraction (HFpEF) with an EF value greater than 50%; and a borderline category called heart failure with mid-range ejection fraction (HFmrEF) having EF between 40 to 49% [1]. Previous studies on changes in EF over time in patients diagnosed with HF were mainly conducted through heart failure registry data or site-specific clinical trials involving limited patient cohorts. The widespread adoption of electronic health records (EHRs) has resulted in a dramatic increase in clinical research using EHR data, especially for phenotypic information and cohort identification. In this study, following a previously validated algorithm [2] for HF phenotype, we identified subphenotypes based on EF values, and studied the longitudinal changes in EF over time using a multi-site retrospective study design.

Methods

This multi-site study was conducted at three large academic medical centers: Weill Cornell Medicine (New York City, NY), Mayo Clinic (Rochester, MN) and Northwestern Medicine (Chicago, IL) following approval from corresponding institutional review boards (IRB). We denote these sites as Site-A, Site-B and Site-C in random order for anonymizing results from individual sites. Echocardiograms from site EHR systems were the source for patients’ EF values. Values were available either natively as structured data or extracted from echo reports through natural language processing. Inclusion criteria for the base patient cohort consists of (1) All patients 18 years or older from 2000-2019, (2) Current Procedural Terminology (CPT) code for echocardiogram, (3) International Classification of Disease-Ninth Revision (ICD-9) codes for diagnosis of HF (any 428.xx) or ICD-10 codes for diagnosis of HF (any 150*), (4) Either B-type natriuretic peptide (BNP) or pro-BNP (NT-proBNP) values recorded, and (5) Prescriptions for HF medication within 6 months of diagnosis. Subphenotypes of HF were identified by the EF value assessed at index date. From the above base HF population, we further identified patients with at least 2 consecutive EF assessments over a 1-year duration. Transitions from HFpEF to HFmrEF, HFpEF to HFrEF, and HFmrEF to HFrEF were pooled and defined as EF-

### Table 1: 1-year metrics for HF patients with EF changes at three sites

<table>
<thead>
<tr>
<th>EF Change</th>
<th>Site-A</th>
<th>Site-B</th>
<th>Site-C</th>
</tr>
</thead>
<tbody>
<tr>
<td>HF patients (n)</td>
<td>10,019</td>
<td>9,270</td>
<td>64,475</td>
</tr>
<tr>
<td>HF patients with 2+ EF measurements</td>
<td>7,841</td>
<td>3,547</td>
<td>28,160</td>
</tr>
<tr>
<td>Age (mean year)</td>
<td>65.5</td>
<td>65.8</td>
<td>67.5</td>
</tr>
<tr>
<td>Female (%)</td>
<td>46</td>
<td>45</td>
<td>35</td>
</tr>
<tr>
<td>EF-Stable</td>
<td>6,242 (79%)</td>
<td>2,356 (67%)</td>
<td>20,413 (73%)</td>
</tr>
<tr>
<td>EF-Decrease</td>
<td>599 (8%)</td>
<td>508 (14%)</td>
<td>3,209 (11%)</td>
</tr>
<tr>
<td>EF-Increase</td>
<td>1,000 (13%)</td>
<td>683 (19%)</td>
<td>4,538 (16%)</td>
</tr>
</tbody>
</table>

1-year Transition among subphenotypes

<table>
<thead>
<tr>
<th>Transition</th>
<th>Site-A</th>
<th>Site-B</th>
<th>Site-C</th>
</tr>
</thead>
<tbody>
<tr>
<td>HFpEF to HFmrEF</td>
<td>184 (4%)</td>
<td>178 (8%)</td>
<td>1,497 (9%)</td>
</tr>
<tr>
<td>HFpEF to HFrEF</td>
<td>246 (6%)</td>
<td>277 (12%)</td>
<td>806 (5%)</td>
</tr>
<tr>
<td>HFmrEF to HFrEF</td>
<td>169 (27%)</td>
<td>53 (20%)</td>
<td>906 (26%)</td>
</tr>
<tr>
<td>HFmrEF to HFpEF</td>
<td>217 (34%)</td>
<td>160 (61%)</td>
<td>1,459 (41%)</td>
</tr>
<tr>
<td>HFpEF to HFmrEF</td>
<td>281 (10%)</td>
<td>94 (9%)</td>
<td>1,452 (18%)</td>
</tr>
<tr>
<td>HFrEF to HFpEF</td>
<td>502 (17%)</td>
<td>429 (44%)</td>
<td>1,627 (21%)</td>
</tr>
</tbody>
</table>
Decrease. Transitions from HFrEF to HFmrEF, HFrEF to HFpEF, and HFmrEF to HFpEF were pooled and defined as EF-Increase. Those patients with no change among EF subphenotypes were pooled and defined as EF-Stable.

Results

Table 1 lists metrics for HF patients at the three participating sites. Figure 1 shows 1-year longitudinal changes in EF values from the corresponding baseline values for HF patients who had at least 2 EF measurements available, across the three sites.

![Figure 1. EF changes in 1 year among HF subphenotypes at three sites. Each bar segment shows the proportion (%) of patients with changes with stable EFs in each category as references.](image)

Discussion

In this multi-site EHR based retrospective study, we sought to investigate the nature of longitudinal EF changes in HF patients treated for medical care at three multi-specialty hospital systems and compare them with previous studies that involve either a small cohort data or previously curated HF registry data. Using an HF registry data, Savarese et. al. [1] studied 4,942 patients (mean age 72 years, 31% female) for their EF changes during a mean follow-up time of 1.4 years, and a baseline EF distribution of 18% (HFpEF), 19% (HFmrEF), and 63% (HFrEF). During follow-up, 21% and 18% of HFpEF patients transitioned to HFmrEF and HFrEF, respectively; 37% and 25% of HFmrEF patients transitioned to HFrEF and HFpEF, respectively; and 16% and 10% of HFrEF patients transitioned to HFmrEF and HFpEF, respectively. These results were substantially different from our current observations at three sites. For all three sites, a smaller percentage of HFpEF patients reduced their EF values to HFmrEF and HFrEF. Similarly, a higher percentage of HFrEF patients improved their EF values to HFmrEF and HFpEF. Among other things, we attribute these differences mainly to patient characteristics at these centers, who are younger and have a higher proportion of females compared to the reported study. These results demonstrate the importance of replication studies to represent geographically distinct populations more broadly. Data gathered from this longitudinal study were further used to develop various machine learning models to predict EF changes in heart failure patients and will be the subject of a forthcoming manuscript.

Conclusion

While we observed differences in EF changes between the three participating sites, the observed results in general varied substantially from a previous study that employed HF registry data.

References


1291
Toward phenotyping of ventilator-induced lung injury with a damage-informed pulmonary model of lung-ventilator interaction.

David Albers1,2,3, PhD, Deepak Agrawal1,3, PhD, Brad Smith3, PhD, Peter Sottile4, MD, Tellen Bennett1, MD MS, Jake Stroh1,3, and George Hripcsak2, MD MS

1Pediatrics, Section of Informatics and Data Science, University of Colorado, Aurora, CO, 2Biomedical Informatics, Columbia University, New York, NY; 3Department of Biomedical Engineering, University of Colorado, Denver, CO, 4Internal Medicine, Division of Pulmonary Sciences and Critical Care, University of Colorado, Aurora, CO.

Introduction
Mechanical ventilation is a life-saving therapy that is challenging to apply because of complex interplay between ventilator mechanics and settings, patient-ventilator interactions, disease processes, non-ventilator interventions (e.g. sedation), and the underlying pulmonary physiology[1], [2]. Mechanical ventilation can worsen lung injury through barotrauma (e.g., alveolar rupture due to pressure differences imposed by mechanical ventilation), volutrauma (e.g., overextension of alveoli), and atelectrauma (e.g., shearing forces from repeated collapse and re-inflation) that are referred to collectively as ventilator induced lung injury (VILI) [1], [2]. Clinicians attempt to minimize VILI but mortality from VILI-related syndromes, e.g., acute respiratory distress syndrome, remains unacceptably high [3].

Identifying ventilator settings that minimize VILI is challenging because the demands of gas exchange and lung protection are frequently in conflict; e.g. low pressures protect from VILI but may not provide sufficient gas exchange. Avoiding VILI is further complicated by the fact that it is difficult to understand and quantify lung injury. Both inter- and intra-patient heterogeneity is the norm. Injury is not homogenously [4] distributed within the lungs of a patient. Furthermore, the condition of the lungs change over time as injury worsens or resolves. Unfortunately, we currently lack the methodological ability to define quantifiable, interpretable phenotypes of lung function anchored to physiologic understanding that could be used to guide ventilator management. In addition to injury solely caused by the ventilator settings, VILI may be instigated and exacerbated by patient interactions with the ventilator, or ventilator dyssynchrony (VD). There are roughly seven types of VD that can occur at varying frequencies, with varying severities, and in varying confounding circumstances[1], [4], [5]. For example, the patient may inspire more than the prescribed tidal volume and cause volutrauma. While it is easy to imagine how VD may cause VILI, e.g., by straining the lung tissue, exact mechanisms and the types of VD that cause the worst injury are not well understood. Specifically: (i) the causal mechanics between types of VD and VILI are not well described, (ii) quantification of observable VILI in bedside ventilator data is missing, and (iii) quantitative understanding—e.g., a rank ordering—of types of VD regarding their level of damage and impact on outcomes does not exist. We do know that VD is strongly associated with mortality and that this association is believed to be causal. However, understanding of how, when and why VD drives VILI and mortality is missing. It has proven difficult to deeply and quantitatively investigate the relation between VD and VILI. To solve this problem and provide raw materials—knowledge extracted from data—for decision support, we need methods that can deduce new knowledge from ventilator data that can be synthesized into understandable phenotypes of degree and kind of lung injury without which clinical decision support is not possible. In this abstract we present a new mechanistic model-based method for estimating lung states from ventilator data. We show how the method can extract lung-damage information from the data not easily extracted otherwise. Similarly, we show how the new method compares to other state of the art lung-modeling efforts.

Methods
We use prospectively collected human ventilator and mouse-model data to validate a newly developed damage-informed lung model (DILM). The DILM is represented as 8 ordinary differential equations modeling pressure and volume, with four constraint equations and 20 parameters. The model, at baseline, models the healthy pressure and volume ventilation waveforms or signals. Damage or lung injury can be conceptualized as deviation from ideal pressure and volume wave forms. While there are many potential ways that an observed pressure or volume waveform can deviate from an ideal, there are particular deviations that are hypothesized to correspond to important types of lung damage. For example, some deviations indicate that a portion of the lung collapsed and only opened once it was forced open by elevated pressure. The DILM takes a healthy baseline and then adds functional terms that correspond to particular, damage specific, deviations, incorporating clinical and physiologic knowledge into the models via allowed pressure and volume diversity. This allows the model to accommodate and estimate deviation of the waveform...
Figure 1: Damage-informed lung model estimates two mouse breadths (top), and two human breadths (bottom) and a comparison of estimation error for single compartment vs. damage-informed lung models for 150 human breadths with mean squared errors of 0.25 and 0.09 respectively (far right).

estimate the same model for both humans and mice in similar situations—types of initial injury, types of VD, etc.—with the end goal of physical validation and explanation of the calculated phenotype by studying the mouse lungs. The preliminary work here includes estimates of 100s of breadths from 2 individuals and 8 mice. Evaluation of model error is quantified using the mean squared error between the data and the model.

Results
Figure 1 shows two different breaths from a mouse (top) and an ICU patient with ARDS (bottom), one healthy breath at the beginning of the experiment or middle of the ICU stay, and one injured breath at the end of the experiment or ICU stay, with MSE Volume<1.0E-5, Pressure<1.0E-2 and Volume<1.0E-3, Pressure<1.0E-1 respectively. The pressure-volume (PV) loops reveal different lung states (phenotypes). For the paralyzed mouse, the observed PV loops indicate a reduction in lung compliance and increase in hysteresis, characteristic of lung injury. The patient was sedated but not paralyzed and the ventilator was in pressure-controlled, patient triggered mode, minimizing patient-ventilator interactions. We observe lung compliance is substantially improved prior to extubation. These observations are quantified into computable phenotypes via the estimated parameters showing the ratio of tidal (peak) volume to peak pressure increased while the PEEP parameter was reduced. Finally, the bottom right panel of Fig. 1 shows the accuracy gain with the DILM versus the single compartment model via MSE for 120 sequential breadths: the damage informed lung model outperforms the single compartment model by a factor of 3 on non-VD breadths.

Conclusion
We developed a new model and used DA methods to estimate explainable and concordant lung phenotypes in mice and humans that differentiated health states using a single breadth. This provides a methodological pathway for personalized knowledge extraction and decision support for mechanical ventilation.

References
Imprecision and Preferences in Interpretation of Verbal Probabilities in Health: A Systematic Review

Katerina Andreadis, MS1 Ethan Chan,PharmD, MS1,2 Minha Park,MS1,2 Natalie C Benda, PhD1 Mohit M Sharma, MPH1 Michelle Demetres, MS3 Diana Delgado, MLS3 Elizabeth Sigworth, BA,4 Qingxia Chen, PhD,4 Andrew Liu, BS4 Lisa Grossman Liu, MPhil4 Marianne Sharko, MD, MS1
Brian J Zikmund-Fisher, PhD5 Jessica S Ancker, MPH, PhD1,7

1Department of Population Health Sciences, Weill Cornell Medicine, New York, NY; 2Memorial Sloan Kettering Cancer Center, New York, NY; 3Samuel J Wood Library, Weill Cornell Medicine, New York, NY; 4Department of Biostatistics, Vanderbilt University Medical Center, Nashville, TN 37203; 5Department of Health Behavior and Health Education, University of Michigan School of Public Health, Ann Arbor, MI; 6Department of Biomedical Informatics, Columbia University Vagelos School of Medicine, New York, NY; 7Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN

Introduction
Making informed health decisions is dependent on the ability to discern and interpret the information relayed. This becomes more challenging in patient-facing technologies that do not provide electronic health record integration, or direct provider communication, placing an even heavier burden on patients to interpret their health data accurately. Although probabilities of risk and benefit are often essential components of medical information, it is well established that many patients have low numeracy, which may impair their ability to understand or make decisions based on numerical information.1-5 As a result, many technology developers are concerned that patients will not be able to use numerical information or prefer words to numbers. Information technology standards emphasize the development of tools for interpreting and improving the quality of data, but there is limited guidance on how to accomplish this. Therefore, this study's objective was to assess patient interpretation and preferences for verbal probability information, such as the terms “rare” or “common”, which could be used to relay health risks and quantities such as laboratory data in patient-facing technologies.

Methods
We conducted a systematic review of the literature published through September 2020 following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA)6 statement to identify studies using verbal probability terms. Original studies conducted in English with samples representative of lay populations were included if they assessed health-related information and elicited either (a) numerical estimates of verbal probability terms and/or (b) preferences for verbal vs. quantitative risk information.

Data extracted included the main question or comparison, the outcomes measured (either numerical interpretation of the probability terms, format preferences, or both) and the population recruited. For the outcome of numerical estimates of a verbal probability, we recorded sample sizes, mean estimates, ranges, and either standard deviations or 95% confidence intervals. In the list of verbal terms, we did not distinguish between adjectival and adverbial forms (e.g., rare and rarely). For format preference, we recorded the proportions of subjects who reported preferring words, numbers, both, other, or no preference. To assess the risk of bias, we adapted criteria from the AHRQ Methods Guide for Comparative Effectiveness Reviews and Cochrane Handbook for Systematic Reviews of Interventions.7,8

Results
We identified 33 original studies that referenced 145 verbal probability terms, of which 45 were included in at least two studies, and 19 in three or more. Of the studies, 15 focused on medication side effects, 14 on disease risks, and 4 with no specified context. In 24 studies, numerical estimates were elicited for verbal probability terms. Table 1 reports weighted averages and ranges for 7 terms that were evaluated in at least four studies each. The term “rare” was estimated to mean an almost 10% risk, whereas the term “likely” averaged 72%. Variability of interpretation of these terms was high, both across studies (minimum and maximum study averages reported in column 5 and 6) and within study (ranges reported in column 7). For example, individuals estimated the term “rare” to mean anything from 0% to
80%, and “common” to mean anywhere between 10% and 100%. Among 2 large studies of European Commission guidance (EC) verbal labels, including 1,053 participants, an average of 70.1% misinterpreted the EC risk label.

Table 1: Numeric estimates of verbal probability terms studied in 4 or more studies

<table>
<thead>
<tr>
<th>Verbal Probability Term</th>
<th>Number of studies</th>
<th>Average numeric estimate, random effects model (%)</th>
<th>95% CI (%)</th>
<th>Minimum sample average (%)*</th>
<th>Maximum sample average (%)*</th>
<th>Range of individual estimates (%)†</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rare(ly)</td>
<td>7</td>
<td>10.00</td>
<td>[7.99, 12.01]</td>
<td>7.0</td>
<td>21</td>
<td>0-80</td>
</tr>
<tr>
<td>Uncommon</td>
<td>4</td>
<td>17.64</td>
<td>[13.19, 22.09]</td>
<td>13.3</td>
<td>22.9</td>
<td>0-90</td>
</tr>
<tr>
<td>Unlikely</td>
<td>6</td>
<td>17.71</td>
<td>[14.86, 20.55]</td>
<td>13.3</td>
<td>27</td>
<td>0-85</td>
</tr>
<tr>
<td>Possible(ly)</td>
<td>6</td>
<td>43.28</td>
<td>[36.66, 49.89]</td>
<td>36.9</td>
<td>62</td>
<td>--</td>
</tr>
<tr>
<td>Common</td>
<td>6</td>
<td>58.73</td>
<td>[50.40, 67.06]</td>
<td>34.2</td>
<td>70.5</td>
<td>10-100</td>
</tr>
<tr>
<td>Probable(ly)</td>
<td>5</td>
<td>69.87</td>
<td>[67.07, 72.67]</td>
<td>66</td>
<td>73.9</td>
<td>20-100</td>
</tr>
<tr>
<td>Likely</td>
<td>6</td>
<td>71.87</td>
<td>[69.90, 73.84]</td>
<td>66</td>
<td>94</td>
<td>--</td>
</tr>
</tbody>
</table>

In 12 studies, participants’ preference for verbal versus numeric information was captured. In 9 of the 12, majorities preferred numeric risk information alone or in combination with verbal labels. In the 5 studies that permitted a choice between verbal, numeric and combined formats, about one-third of respondents preferred the combination of numeric with verbal descriptions.

Discussion
Since 1967, 33 studies have examined lay interpretation of and preferences for verbal probability terms in health contexts. These studies show numeric interpretations of verbal terms are extremely variable and highly overlapping, providing no assurance that patients will perceive a health outcome described as “common” as more likely than one described as “rare.” The literature suggests that patients prefer numeric risk information, alone or in combination with verbal labels, therefore designing patient-facing technologies that use verbal-only risk descriptors may not be meeting patients’ preferences, resulting in poor communication. A bigger emphasis should be placed on evaluating numerical representation formats of health data and developing evidence-based guidelines on communicating probabilistic health information to empower patients to make their best-informed medical decisions.

Acknowledgments: This project was supported by NLM R01 LM012964 (Ancker, PI).

References
Spillover Effects from the HITECH Act on Innovation in Medical Informatics

Nate C. Apathy, PhD¹, A Jay Holmgren, PhD², Shane Greenstein, PhD³
¹University of Pennsylvania, Philadelphia, PA; ²University of California – San Francisco, San Francisco, CA; ³Harvard Business School, Cambridge MA

Introduction

The Health Information Technology for Economic and Clinical Health (HITECH) Act of 2009 stimulated the United States economy and digitized the health care delivery system via generous subsidies.¹ Beyond this first-order goal, the HITECH Act was projected to generate an important second-order effect: an ecosystem for digital health innovation built on digitized clinical data and broad adoption of information technology systems among provider organizations.² While some progress has been made, broader economic indicators of the growth of the industry are less well-understood. Furthermore, the breakdown of how any returns were distributed across geographies, incumbent firms, academic innovators,³ and major EHR vendors is largely unknown, leaving the winners and losers of the HITECH Act unclear. In this study, we analyzed the second-order innovation returns to the HITECH Act in the form of medical informatics patenting activity to determine HITECH’s overall impact on innovation as well as any distributional effects across a variety of dimensions.

Methods

We used detailed patent-level data from two databases spanning the US and European patent offices from 2001 to 2017. For our first research question examining the impact of the HITECH Act on overall innovative activity in medical informatics, we compared medical informatics patenting in the US to Europe using a difference-in-differences study design, leveraging the PATSTAT database of global patent applications. For our subsequent research questions analyzing the distribution of innovation returns within the US, we compared US medical informatics patents to US non-medical informatics patents using the USPTO PatentsView database, which includes information on all granted US patents. We analyzed only granted patents, and stopped our data as of 2017 to account for multi-year patent approval timelines. We identified medical informatics via the PATSTAT International Patent Classification or PatentsView Cooperative Patent Classification (CPC) Subclass of G16H, Healthcare informatics (i.e., information and communication technology specially adapted for the handling or processing of medical or healthcare data). Informatics patents for our non-medical comparison group were identified as CPC Subclass G06Q, Data processing systems or methods specially adapted for managing, promoting or practicing commercial or financial activities.

In our analyses, we used ordinary least squares multivariate regression models and a difference-in-differences specification \( Y = \beta_0 + \beta_1\text{TreatmentGroup} + \beta_2\text{PostHITECH} + \beta_3\text{TreatmentGroup*PostHITECH} + e \) to estimate the causal impact \( (\beta_3) \) of the HITECH Act on 1) overall innovative activity in medical informatics, 2) innovative activity in new economic areas, and 3) innovative activity among new firms. We used OLS multivariate regression models and an interrupted time series (ITS) specification \( Y = \beta_0 + \beta_1\text{StudyMonth} + \beta_2\text{PostHITECH} + \beta_3\text{StudyMonth*PostHITECH} + e \) to estimate the impact of the HITECH Act \( (\beta_3) \) on the rate of change in innovative activity among 4) academic organizations and 5) major EHR vendors. For analysis (1), we made two comparisons. First, for our international comparison, our outcome \( Y \) is the logged number of monthly medical informatics patent applications and the treatment group is US patent applications. Second, for our within-US comparison, our outcome \( Y \) is the logged number of monthly medical informatics patents and the treatment group is US G16H patent applications (vs. US G06Q applications). For both models, the post period is defined as Feb. 2009 through Dec. 2017. Analyses (2) through (5) use outcomes defined as aggregated monthly proportions of patents attributable to the dimension of interest (e.g., % of patent applications in new economic areas). New economic areas are defined as BLS Economic Areas that were not in the top 10 most common medical informatics areas prior to HITECH. New firm entrants are defined at the firm-year level, with firms identified as “new entrants” if they have no previous medical informatics patent applications. Academic assignees were defined as those with a university or academic medical center affiliation (e.g., Regents of the University of Texas or the Mayo Clinic). Finally, major EHR vendors were defined as any of the top 10 ambulatory or inpatient EHR vendors as of 2017, per the Office of the National Coordinator for Health IT.
Results

Overall innovative activity in the US medical informatics industry increased in the years following the HITECH Act’s passage. From 2001-2008, the US averaged 602 annual medical informatics patent applications, compared to the 2009-2017 period, which averaged 1,338 annual applications, a 122% increase (Figure 1). In difference-in-differences analyses comparing the rate of innovative activity growth in the US to Europe, we found a 68.0% greater rate of growth in medical informatics innovative activity in the US as a result of the HITECH Act ($\beta$=0.519, $p$=0.001). Within-US estimates for the law’s effect on innovative activity were similar (41.1%, $\beta$=0.344, $p$=0.001). In our within-US comparisons to non-medical informatics patenting, the HITECH Act was responsible for an increase of 4.1 percentage points in the share of innovative activity in new economic areas ($\beta$=4.10, $p$<0.001), but did not have an effect on new firm share of innovative activity ($\beta$=2.54, $p$=0.07). We found more new firms annually on average after HITECH – 39% more than pre-HITECH – but new firms’ share of innovative activity did not increase as a result of the law. In interrupted time series analyses comparing the rate of growth in academic and major EHR vendor patenting, we found that the HITECH Act had a positive effect on the share of academic patenting ($\beta$=0.034, $p$=0.02), but did not affect vendor innovation ($\beta$=0.028, $p$=0.09). These two groups accounted for 2.1% -7.8% annually of all medical informatics innovation during the study period (Figure 1).

Discussion

We found substantial returns to innovative activity in medical informatics following the passage of the HITECH Act. While the HITECH Act led to major increases in US medical informatics innovation, our distributional findings of those innovation returns are noteworthy. First, rather than exacerbating geographic disparities in innovation activity, the HITECH Act’s innovation returns accrued in non-traditional technological hubs. This may be in part due to activity from academic medical centers and the local nature of health care markets. Second, while new economic areas were brought into the fold as well as a large number of new firms, the share of innovation from new firms did not change as a result of HITECH. This may be due to the high barriers to market entry in medical informatics, including high regulatory burdens put in place by HITECH itself. Third, we do not find a cooling effect among academic informatics innovators, a concern raised by the informatics community as more academic centers have transitioned to vendor EHR systems. On the contrary, our results imply that as a share of overall patenting, academic innovators have seen an increase in innovation since HITECH. Finally, the large infusion of cash to EHR vendor balance sheets does not appear to have resulted in substantive increases in innovation from these organizations. As vendors have increasingly needed to provide support and maintenance services, the influx of EHR adoption driven by HITECH may have crowded out development activities, a concern given the substantial innovations required to improve EHR usability.

Conclusion

The HITECH Act’s adoption incentives had a large, positive second-order effect on innovation in medical informatics. The distributional composition of those effects suggest that future innovation needed for usability improvement may be more likely to come from non-traditional innovation centers and academic innovators, rather than EHR vendors.

References

A Generalizable Framework for Cost-Effectiveness Analysis of Antihypertensive Drugs Leveraging Real-World Evidence

D. Arneson, PhD¹, R. Vashisht, PhD¹, V. A. Rudrapatna, MD PhD¹, A. J. Butte, MD PhD¹
¹University of California, San Francisco, CA

Introduction
Hypertension is a chronic disease that affects one in three adults in the United States with an estimated annual direct cost of $47.3 billion and annual patient cost of $733¹. Despite the widespread incidence and financial burden of hypertension, there does not exist a standardized set of prescription guidelines. What exists is a multitude of guidelines from different countries, agencies, and associations with varying recommendations spanning multiple drug classes and active ingredients leading to prescription tendencies which may not be optimized for cost-effectiveness.

To bypass the technical limitations imposed by the reanalysis of randomized controlled trials (RCTs), we capitalized on the vast and growing wealth of prescription drug data residing in electronic health records (EHRs). These data were the basis of our novel generalizable framework to conduct cost-effectiveness analysis of prescription drugs using real-world evidence (RWE). This framework allowed us to directly compare the cost-effectiveness of all antihypertensives to propose a new set of guidelines as ranked lists of effectiveness per dollar based on sets predefined factors.

Methods
For a given drug class, we obtained the deidentified EHR records of all patients who were treated with that drug class and met a prespecified set of inclusion criteria. The inclusion criteria for this study were: first time user of antihypertensive monotherapy, prior diagnosis of hypertension with blood pressure (BP) ≥ 130/80, prescription occurred during an outpatient visit, at least 1 year of medical history in the EHR prior to the prescription, at least 1 year of follow-up in the EHR after the prescription, each consisting of at least two visits with a physician with primary care as their primary specialty, and at least 60 days of consecutive use of the prescription.

To evaluate the cost-effectiveness of antihypertensives on patients commonly seen in the health care system, cohorts were specified for exhaustive combinations of demographics (age, sex, race, etc.) and comorbidities (type 2 diabetes, chronic kidney disease, heart failure, etc.). The resulting cohorts were 1:1 propensity score matched for all pairwise combinations of antihypertensive drug classes. We considered a number of outcomes to evaluate the relative cost-effectiveness of each drug including quality-adjusted life year (QALY), incremental cost-effectiveness ratio (ICER), ΔBP per dollar, and incidence of secondary outcomes per dollar (e.g. heart failure, stroke, dementia, etc.). QALY and ICER utility values included days as inpatient/outpatient, acute stroke, chronic stroke, myocardial infarction, acute unstable angina, coronary heart disease, and death and were assessed at 1-, 3-, and 5-years post-treatment. The follow-up for other outcomes was at the first incidence of: outcome, 5 years follow-up, 12-31-2020, discharged dead, and stopped or changed antihypertensive. Based on the selected outcome, drugs were ranked by benefit per dollar using national average drug pricing obtained from the IBM Micromedex RED BOOK.

We first built our framework using EHR data from the UCSF OMOP Deidentified (DEID) Clinical Data Warehouse (CDW) which uses the OMOP Common Data Model (CDM). This standardized data format allowed us to validate our findings using EHR data from four other academic medical centers which are part of the University of California Health system (UC Davis, UCLA, UC Irvine, and UCSD). The UC Health System EHR is ~5x the size of the UCSF EHR which allowed us to evaluate an additional ~30,000 patients who met our inclusion criteria.

Data was queried from the UCSF OMOP DEID CDW using SQL and from the UC Health Data Warehouse (UCHDW) through the UC Data Discovery Portal with Apache Spark and Databricks. All downstream analysis was conducted using R statistical software. Reported p-values have been adjusted for multiple testing using Bonferroni correction. Code will be made available to facilitate reproducibility and dissemination of the framework.

Results
We first applied our framework to 3,203 antihypertensive treated patients from the UCSF OMOP DEID CDW who met our inclusion criteria. To identify the antihypertensive drug class which produced the greatest change in blood pressure per dollar, we considered all pairwise combinations of drug classes and matched the cohorts with 1:1 propensity score matching to adjust for potential bias.
To evaluate the generalizability of both our framework and these findings, we extended our analysis to the UCHDW which included four additional academic medical centers (Figure 1). We observed a similar effect on change in systolic blood pressure between diuretics and three of the four other antihypertensive drug classes, with few statistically significant differences (Table 1 top). However, ACE inhibitors provided a greater reduction in systolic blood pressure across all five sites with a significant difference at UCLA and close to significant at UCD, UCI, and UCSF (adjusted $p$-value = 0.056, 0.05, and 0.068) compared to a matched cohort of diuretics users. These absolute changes in systolic blood pressure were then normalized to cost per day for a 10 mm Hg decrease in systolic blood pressure using cost and frequency information (Table 1 bottom). Diuretics provided the lowest cost per day for a 10 mm Hg reduction in systolic blood pressure (range of average values: $0.10$-$0.18$) versus the other four reported antihypertensive drug classes (range of average values: $0.58$-$1.63$). This is largely attributable to the similar effect of diuretics on the outcome of change in systolic blood pressure versus other antihypertensive drug classes and the low average cost of diuretics versus the average cost of the other drug classes.

**Table 1.** Results of cost-effectiveness of antihypertensives from UCHDW. Pair-wise comparisons for drug class (each row) versus diuretics across 5 sites in the UC Health System (columns). The top half of the table reports changes in systolic blood pressure after antihypertensive treatment and the bottom half reports the normalized price for a 10 mm Hg change in systolic blood pressure. For each site, the two columns give the average change in systolic blood pressure (cost per 10 mm Hg decrease) with the left number corresponding to diuretics and the right number corresponding to the drug class indicated in the row. Bolded numbers have adjusted $p$-value $< 0.05$ from a Wilcoxon rank sum test between matched cohorts of diuretics and the indicated drug class. ARBs (Angiotensin Receptor Blockers), CCBs (Calcium Channel Blockers).

**Discussion**

We developed a flexible framework that can be used to evaluate the cost-effectiveness of all drugs for a given indication. This framework was used to systematically evaluate the benefit-per-dollar of all commonly prescribed antihypertensives using RWE. Our findings were robust across multiple hospital centers in the UC Health System. The outcomes of this study can be used to inform new prescription guidelines for patient populations treated with antihypertensives. While cost-effectiveness is an important consideration in the medical decision-making process, other factors which may not be modeled by our framework also contribute (e.g. patient insurance status, side effects, medication adherence, etc.). The framework we introduce here highlights a new use for RWE that can make an immediate impact on clinical prescription guidelines for any drug class it is applied to.

**References**

Design for Health System Resilience in Challenging Times: 
A Framework for Remote Cancer Care Through Community Codesign

Eliah Aronoff-Spencer, MD PhD¹, Melanie McComsey, PhD¹; Alexandra Hubenko, MBA¹, Mingyuan Chih, PhD², Timothy W. Mullett, MD MBA FACS², Corey E. Baker, PhD², John Kim, EdD MPH³, David K. Ahern, PhD³, Michael C. Gibbons, MD MPH⁴, Joseph Cafazzo, PhD⁵, Pia Nyakairu, BID⁵, Robin C. Vanderpool, DrPH⁶, Bradford W. Hesse, PhD⁷  
¹University of California San Diego, La Jolla, CA; ²University of Kentucky, Lexington, KY; ³Brigham and Women’s Hospital, Boston, MA; ⁴Johns Hopkins University, Federal Communications Commission (FCC) Connect2Health Taskforce, Washington, DC; Baltimore, MD; ⁵University Health Network, Toronto; ⁶National Cancer Institute, Bethesda, MD; ⁷Retired (National Cancer Institute), Kailua-Kona, HI

Introduction
During the COVID-19 pandemic, screening and diagnostic testing have been significantly curtailed or delayed, resulting in projections of a substantial rise in excess deaths from cancer¹. An opportunity exists to redesign systems of cancer care in a virtual world that are more inclusive of stakeholder input and resilient through participatory design, community engagement, and iterative collaborative development of robust and sustainable solutions. This study describes a human-centered, participatory design approach to engage stakeholders in solving a community-defined problem: here, the problem of serving rural patients experiencing distress during cancer treatment. The effort represents the work of a public-private partnership referred to as the Linking and Amplifying User-centered Networks through Connected Health, or the L.A.U.N.C.H. Collaborative². In L.A.U.N.C.H., priority areas are framed where stark differences in cancer burden and resources present an opportunity for innovation. Data from the national cancer registry and Federal Communications Commission has pinpointed Appalachian Kentucky as an area experiencing a “double burden” of poor cancer outcomes and lack of access to broadband. It is here that we focused the L.A.U.N.C.H. work with the hypothesis that codesign of solutions prioritized by, and built with diverse local stakeholders, would be more effective, scalable and resilient than traditional approaches. We describe the codesign and co-creation processes in developing the MyPath app for remote cancer distress management in Appalachian Kentucky and present results from formative evaluation of the new service model as compared to the gold standard, paper-based cancer distress monitoring tool, the National Comprehensive Cancer Network Distress Thermometer (DT).

Methods

People, Context and Design. Following the Institutional Review Board approval, cancer patients, family caregivers, providers, payers, technologists and broadband providers were engaged across Appalachian Kentucky and within the University of Kentucky Markey Cancer Center. A formal design process with roots in Human Centered Design for Complex Sociotechnical Systems³⁴ was adopted. We developed the “L.A.U.N.C.H. ROADMAP” as a guiding framework of both innovative and proven approaches. The overall approach is outlined in prior work²⁵.

Designing, Prototyping and Testing. The L.A.U.N.C.H. ROADMAP contains seven Phases: Identify, Discover, Define, Ideate, Refine, Implement, and Test. In the Identify phase, existing data was analyzed, and stakeholders provided input to identify cancer distress management in Appalachian Kentucky as the key issue. In Phase 2, the stakeholders were engaged in Kentucky using ethnographic fieldwork to understand the cancer experience and to develop a network of local stakeholders for subsequent codesign activities. With stakeholder input, the problem was redefined in Phase 3. In Phase 4, ideation sessions were introduced to help participants constrain solutions but allowed for local innovation based on resources that had not been considered prior. In Phase 5 and 6 (the 72-hour co-design sprint), prototypes were described verbally, then refined with low-fidelity methods such as post-it notes. In parallel, a professional designer (Toronto, Canada), a mobile application developer (Berlin, Germany) and a design support group (San Diego) worked together to translate low-fidelity concepts to working digital applications. These working prototypes were tested and iterated in real time with participants in Kentucky. In Phase 7, two booths in Kentucky were staged to collect feedback about the codesigned instrument (MyPath), as compared to the DT. Participants were introduced to the paper version of the two instruments first, then followed by the app version. Providers were invited to evaluate the digital dashboard. Participants were given an opportunity to use these prototypes and asked to complete a 10-item System Usability Scale (SUS) to evaluate the respective prototypes. Participants and stakeholders in all
these activities are cancer patients, caregivers, providers, and other community stakeholders, including broadband Internet providers and healthcare payers.

Results

**Design Activities.** Problem Identification, contextually Discovery and problem Refinement set the stage for participatory design work. The Ideation phase was initiated with a codesign booth where participants stop and go in Pikeville and the 1st design studio where participants sit and work in Lexington, KY, to frame problems in cancer care and conceptualize new solutions. This work contextualized the 2nd design studio in an Appalachian Kentucky county (Jackson) where community members, patients and caregivers were encouraged to brainstorm granular solutions to problems or make practical those solutions that had been considered. These concepts were presented in a subsequent 72-hour co-design sprint in Lexington to select a lead candidate. A prototyped was created and Refined for the concept of cancer distress management and was rapidly Implemented as an intervention to test with patients, providers, caregivers, and others. The outcome of this work was a new software application and service model, call MyPath, developed with participants and delivered to stakeholders for implementation and formal testing.

**Prototypes and Testing.** As the existing DT is paper based, in our 3rd design studio (Oct 7-8, 2019), a comparable paper tool based on codesign and co-creation work from prior studios was developed first. Key findings from these sessions produced the following requirements (and potential pitfalls of the current approach): (a) compelling user experience; (b) framing the problem from patient perspective (my wellness and my path vs distress thermometer); (c) framing questions positively; (d) adding patient centered questions; (e) giving feedback and actionable information upon completion of the survey. Certain feedback from the design studios could not be implemented using paper-based methods alone and required digitization, e.g., to get feedback based on answers, and to communicate patient distress with providers in a timely manner. On the second day (Oct 8, 2019), participants designed a digital version of the paper app (named MyPath by the participants) as well as a digital version of the DT to assess incremental improvement from added functionality, but also to discern the role of digitization alone. Highlighted user requirements for MyPath included: compelling colors, content from the paper app, new summary screen giving immediate feedback upon completion, and actionable insights with instructions and connection to providers. A provider-centric dashboard was prototyped and designed to integrate with user generated data in the MyPath app.

**Prototype Testing.** On October 9-10, 2019, 38 participants (5 patients, 8 caregivers, 11 providers and 14 others) evaluated the paper version of MyPath and DT, 27 participants (7 patients, 4 caregivers, 8 providers and 8 others) evaluated the app version, and six evaluated the MyPath dashboard. Among instruments, the MyPath app had the highest mean SUS score (82), and followed by the DT app (78), the DT paper (74) and the MyPath paper (73). The MyPath app had a higher usability subscale score than the gold standard, the DT paper (t(63)=2.611, p=0.01) and a better overall SUS score than the MyPath paper (t(63)=2.35, p=0.02). Providers rated all instruments higher when compared to how patients rated. However, patient-rated scores increased through our iterative development path from the DT paper, to a reworded version of the MyPath paper, to the DT app, and finally to the MyPath app. Providers, on the other hand, appeared to be less satisfied with the earlier iteration of the MyPath paper than patients did. This finding may be related to clinical staff having more familiarity with the traditional language in the DT. Nevertheless, providers responded with higher usability and learnability ratings than patients did when exposed to the final MyPath app. Providers also highly rated the MyPath dashboard (mean SUS=90). It appears clinical staff grew more appreciative of the form and intent of the project the closer it progressed to its final operable format.

Conclusion

A human-centered codesign approach such as described here can address the global need for rapid health system innovation that generates effective and locally sustainable solutions, and which can scale laterally to promote resilience in the system. A pilot study is currently testing MyPath as a remote cancer monitoring tool in Kentucky.

References

4. LAUNCH Health UC San Diego. LAUNCH Pad [Internet]. Available from: http://launchhealth.ucsd.edu/launchpad/.
Accurately and Privately Reporting Crowdsensed COVID-19 Data

Hafiz Asif, Ph.D., Periklis A. Papakonstantinou Ph.D., Stephanie Shiau, Ph.D., Vivek Singh, Ph.D., Jaideep Vaidya, Ph.D.
Rutgers University, New Jersey

Introduction

In the early months of the COVID-19 pandemic, testing for COVID-19 remained insufficient. This resulted in a lack of crucial information necessary to track the spread of the disease. To fill this information gap, many symptom-tracking apps were developed to collect self-reported health and demographic data – which is both personal and sensitive – and generate insights about the pandemic, e.g., current and emerging hotspots. However, these apps have two major limitations: (i) the insights are released for prefixed regions (e.g., counties) and do not allow reporting for smaller localities (e.g., neighborhoods), where outbreaks typically occur; or for regions spanning administrative boundaries; (ii) though insights are released as aggregates, no privacy is guaranteed for the people reporting data.

To address these issues, we propose a data-driven spatiotemporal partitioning based approach to develop privacy-preserving symptom-tracking (or crowdsensing) apps. Our approach provides the ability to answer, with a provable privacy guarantee, any number of arbitrarily and adaptively chosen spatiotemporal range (SR) queries to gather insights about the spread of COVID-19, without sacrificing utility. In this context an SR query asks for the number of data points in a given region and a time period, e.g., “How many people reported experiencing COVID-19 symptoms in NYC in the past 14 days?” is an SR query. The method was implemented and deployed via COVID Nearby platform.

Method

The basic idea is to create differentially private (DP) spatially and temporally indexed partitions of the space (e.g., the USA) using spatially and temporally partitioned data, and then compute the DP count of the reports (e.g., the number of symptomatic people) in each partition. These indexed partitions with their corresponding DP counts are used to answer SR queries. While it may appear to be trivial to answer queries (i.e., provide DP counts), in actuality it is significantly complex to create the right partitions to provide both privacy and utility given the arbitrary and adaptive nature of the queries. Note that here for any SR query, at any given day, we consider the past 14 days as its time period.

We propose a hybrid of data-agnostic and data-dependent approaches (in isolation, neither can simultaneously guarantee reasonable privacy and accuracy). First, we partition the space data-agnostically, based on administrative units, e.g., counties – call them divisions. Then, every day for each division, we temporally partition the data of the past 14 days (reported from the division) into k groups (i.e., partitions), and use them to create k different data-dependent partitions of the division by building k-many (ε/n)-DP quadtrees over the same division. Each of these k quadtrees uses the data from one of the k groups and is only built if it has not been built before (here k = ⌈14/n⌉ and n is chosen such that 1 ≤ n ≤ 7). We develop a novel data-partitioning (da-parti) algorithm to create the temporal partitions (every day), ensuring that no report will be included in more than n groups created over time. Thus, the overall privacy risk remains at most ε. To choose n over time, we give an empirically supported heuristic. We also give a heuristic to limit the privacy risk incurred by multiple reports by a user.

To answer any SR query on any given day, we first find all the divisions overlapping the query region. Then, for each of these divisions, we use the k quadtrees (identified by da-parti algorithm) to compute the partial answers, which are aggregated to give the final answer. Since quadtree-based spatial partitioning works for rectangular regions, it introduces some distortion for many of the divisions (i.e., counties). Therefore, we also consider the actual overlapping region (computed via polygonal approximation) of the division to better approximate the answers.

Results

Our method has been validated via an extensive empirical evaluation over spatially disaggregated real-data of confirmed COVID-19 cases. The results show that the DP answers, computed via our method, are highly accurate (Fig. 1(a), (b), (d), (e)). This is true even for the arbitrarily picked region within a division (Fig. 1(b)). Note that our approach yields a much lower error than a baseline approach that provides the same DP guarantee (Fig. 1(c)). Additionally, we can calculate moving averages (Fig. 1(d)) and rank regions very accurately (Fig. 1(e)). We used ε = 6.
Figure 1: Private counts are DP answers to SR queries computed by our method from the actual COVID-19 case count. (a) depicts the stacked bar-chart of the case counts of the 5 NY counties with the most cases. For each day: (i) two stacked bars are given, the first for the actual counts, and the second for the private counts; and (ii) each bar gives the total COVID-19 cases for the 14 days period leading up to that day. (b) juxtaposes the heatmaps of the actual and private 14 day case counts (on a log scale) for NY counties and regions within Richmond county, on 3/20/2020 and 4/13/2020. (c) compares our approach to a naïve approach (i.e., DP data aggregation over partitions created by a fixed grid with cell size of 1 Km²) guaranteeing the same level of privacy. Since both the methods are probabilistic, the boxplots are computed over 100 iterations. (d) plots the 14 days moving average for NY, TX, and the USA. (e) plots the ranking score (i.e., Kendall rank correlation coefficient) in terms of the cases in the past 14 days (i.e., SR queries) for all counties (score = 1 means the data from DP SR queries produces the same ranking as the original data).

Discussion & Conclusions

To use our approach in a similar future pandemic/epidemic, one needs to find the corresponding heuristic to select the parameter \( n \) (for the da-parti algorithm). This can be done by generating synthetic data for the new daily cases by, for example, using SIR model⁵, estimating \( d \) (which for COVID-19 is 14 days), and then following our approach.

The proposed approach can be used to create crowdsensing platforms that guarantee privacy and allow queries across arbitrary space and time bounds. Lack of privacy guarantees has been cited as a leading cause of concern by experts and non-governmental organizations. Hence, the proposed approach can be vital to allaying the concerns of experts and end-users alike. Its support for tracking across administrative boundaries is almost cognizant of the ground realities of the pandemic. Also, the approach can be generalized and applied for crowdsensing information about other health symptoms or adoption behaviors (e.g., vaccination rates). On the whole, this approach paves a way forward for countering pandemics without compromising individual privacy.

Acknowledgments. Research reported in this publication was supported by the NSF under award CNS-2027789 and by the NIH under award R35GM5134927. The content is solely the responsibility of the authors and does not necessarily represent the official views of the agencies funding the research.

References

Applying Real World Data to CDISC Standards through the Joint FHIR® to CDISC® Mapping Document

Rebecca Baker, MS, MHA, BSN, RN1, Rhonda Facile, MS1, Bess LeRoy, MPH1
1CDISC, Austin, Texas, USA

Introduction

As healthcare data has increasingly become digitized, there has been a desire in the research community to harness this information to enhance and assist data collected for clinical research. [1] Clinical Data Interchange Standards Consortium (CDISC®) has a complete suite of standards supporting clinical and non-clinical research processes from planning, data collection, tabulation and analysis [2] all supported by controlled terminology maintained by the National Institute of Health (NIH) National Cancer Institute (NCI) Enterprise Vocabulary Services (EVS). [3] The HL7® Fast Healthcare Interoperability Resources (FHIR®) is a relatively new standard for exchanging healthcare information electronically. The Joint HL7 FHIR to CDISC Mapping IG leverages prior work started by the EHR to CDASH team (E2C) and the TransCelerate Biopharma Inc. to move the needle on achieving greater interoperability and exchange of data from EHRs to clinical research submission-ready datasets. Using FHIR to harness the data for clinical research using CDISC standards has been proposed as a way to capture ‘real world data’ (RWD) where clinical data not directly captured for clinical trial purposes can be used to support regulatory applications. This solution, the Joint FHIR to CDISC Mapping IG, [4, 5], is a step towards making it easier to harness RWD data and leverage it for case report forms (CRFs) and direct capture of data within systems rather than having separate data entry occur in each system.

Methods

CDISC reignited the E2C team to further define the mappings that were created when FHIR was at an early maturity level. A gap analysis was performed on the previous work for completeness and adherence to the current FHIR resources for the SDTM domains of interest: Adverse Events (AE), Concomitant Medications (CM), Demographics (DM), Laboratory (LB), Medical History (MH), Procedures (PR) and Vital Signs (VS). The focus was on FHIR resources with a maturity level of 4 and 5. Gaps were identified and FHIRPath was used to further the mapping. FHIRPath is a path based navigation and extraction language used to express terms of logical content of hierarchical data models in a platform-independent, model-independent expression. A team of FHIR experts did the bulk of the mapping work with insights and direction provided by the E2C team. The mapping was based on FHIR R4 and the latest CDASH version (CDASH IG v2.1, which corresponds to SDTMIG v3.2, SDTM v1.4).

The E2C team performed the reviews in addition to the CDISC community during an Internal Review (IR) process. Once the draft was ready, a simultaneous CDISC Public Review (PR) and HL7 Ballot Review Process occurred with the support of the HL7 Biomedical Research and Regulation Work Group. Feedback was provided from both the CDISC and HL7 communities through this mechanism. Initially the mapping was maintained in an Excel spreadsheet. Once ready for the Public Review, the content was placed in the FHIR GitHub in XML format enabling different viewing formats. This became the “source of truth” for the mapping document. The aim is to publish both the HL7-FHIR IG and the CDISC IG simultaneously as the “Joint FHIR to CDISC Mapping IG”. The publication will include a PDF rendered version of the FHIR IG with supporting mappings and FHIR IG XML source to facilitate loading into the CDISC Library.

Results

There were 153 comments received during the CDISC IR, and a total of 95 comments as part of the joint CDISC and HL7 Public Review process (37 from the CDISC PR and 58 as part of the HL7 ballot process). Once comment resolution is complete, the document will be published simultaneously through CDISC and HL7. The mapping was done at a high level using a vendor agnostic approach with FHIRPath. The mapping deliverable will provide an informative FHIR IG, supplemental mapping document with associated references. A machine-readable version also will be included in the CDISC Library.
Discussion

The community expressed enthusiasm for the project and there were many comments about mapping explicitly end-to-end to promote optimal interoperability.

The rationale for using FHIRPath was that each health system/hospital/clinic has different data elements available in their EHR system and the IT queue in health systems has historically been quite lengthy. Another factor is that each clinical trial has different data requirements that may influence what is collected. As implementers use this document, they will need to be aware of the data they are mapping to and from and the idiosyncrasies within each system.

Challenges identified were (a) the limited pool of experts’ proficient in both FHIR and CDISC standards, (b) limitations of the comprehensive view of the clinical research workflow and (c) a desire expressed to have explicit instructions for the mapping to ensure interoperability. First, there is a small minority of informatics experts with both CDISC standards knowledge and FHIR knowledge. Having a robust review was difficult because expertise was either stronger on one side or the other. Second, understanding the clinical research workflow is important. The data entry workflow means how the data flows from case report form (CRF) to an electronic data capture (EDC) system then submitted for regulatory approval. Team members questioned what the regulation would look like around the mapping and how the data could potentially be traced from entry point to submission. “When does the mapping occur?” and “what are the implications to the end result?” were concerns voiced by some team members. The third challenge noted was the ongoing request for detailed mapping. The community expressed a desire throughout the review process to have a more explicit solution and a more granular approach. The impediment to this is the current ecosystem where each health system has different data elements available for mapping. [6] Future work to make the mapping more explicit would depend on the alignment of healthcare systems worldwide in general to a more consistent standard. This explorative project served to highlight the challenges while also providing a general approach for mapping HL7-FHIR resources to CDISC standards to create compliant datasets, and mapping EHRs to EDC systems.

Conclusion

Real world data can enhance clinical research data by using the HL7-FHIR and CDISC standards. The Joint FHIR to CDISC Mapping IG provides a high-level vendor agnostic approach for implementers to map electronic data to formats used by the clinical research community for submission. Use of this mapping document can support the interoperability of clinical research data from RWD and decrease redundant data entry processes.

References

Toward Patient-Centered Informatics Solutions: The Role of Intersectionality

Emily Bascom¹ & Reggie Casanova-Perez, MS¹, Harshini Ramaswamy², Deepthi Mohanraj¹, Janice Sabin, PhD, MSW¹, Wanda Pratt, PhD¹, Andrea L. Hartzler, PhD¹
¹University of Washington, Seattle, Washington; ²University of California San Diego, San Diego, California

Introduction
Before informatics systems can support equitable, diverse, and inclusive healthcare, we need to understand the impact of intersectionality on people’s experiences. Intersectionality is “the complex, cumulative way in which the effects of multiple forms of discrimination (such as racism, sexism, and classism) combine, overlap, or intersect especially in the experiences of marginalized individuals or groups.” Thus, experienced oppression is not based on a single axis of identity but rather on the interwoven aspects of identities. Coined by Kimberlé Crenshaw in 1989, the term originally referred to the unique oppression that the intersection of race and gender brought to individuals. Today, the term has expanded to include other identities (e.g., ethnicity, gender identity, sexual orientation, religion, body weight, height). Incorporating the concept of intersectionality in the clinical environment, including clinical information systems, acknowledges the whole person and the complexity of patient-provider interactions. This helps to dismantle the cisnormative privileges that are present in the healthcare system, which assumes things such as a person’s gender identity is their sex assigned at birth (i.e., cisgender) and that they are sexually attracted to people from the opposite gender (i.e., heterosexual). These privileges inform providers’ implicit biases, which can contribute to unjust medical treatment of patients with historically marginalized intersectional identities and is perpetuated by health information technology. In this podium abstract, we will present the intersectional identities that Black, Indigenous, People of Color (BIPOC), and/or Lesbian, Gay, Bisexual, Transgender or gender diverse (LGBTQ+) patients shared in their stories related to unfair treatment, elaborate on what we learned from those stories, and discuss the importance of intersectionality for developing solutions to enhance patient-centered care.

Methods
To investigate the lived experience of bias towards race, sexual orientation and gender identity experienced by patients from marginalized groups, we collected twenty-five stories from patients who are BIPOC, and/or LGBTQ+. Patients shared stories in which they not only experienced bias towards their race, sexual orientation, or gender identity, but also simultaneously toward their body weight, age, or height.

Results
We interviewed twenty-five BIPOC and LGBTQ+ participants, ages 19-60 to investigate how they experience biases in healthcare. Eleven participants were BIPOC, three were LGBTQ+, and ten were BIPOC and LGBTQ+. One participant did not report being BIPOC or LGBTQ+ but described themself as an “older Asian woman.” As we expected, seventeen participants brought up racism, seven participants shared discrimination based on sexual orientation, and four participants shared discrimination based on gender identity throughout the stories they shared. Half of the participants (13/25) also shared other types of discrimination; six mentioned sexism (i.e., assumptions made based on biological sex), four mentioned sizeism (i.e., assumptions made based on the person’s weight), three mentioned ageism (i.e., assumptions made based on the person’s age), one mentioned discrimination based on their height, and one mentioned discrimination based on legal, recreational drug use. For example, PT12 explained, “I get mistaken for a high school student or an intern a lot of times at my workplace and it's a struggle... I’m not sure if that's something that's coming slowly from a patient-doctor dynamic or; you know, it's just physical attributes like skin color and height and race that lead to such biases.” These findings provide insight into how the healthcare system fails to adequately and fairly treat the diverse health and wellness needs of patients with intersectional identities. Understanding how intersectional identities play a role in unfair treatment experiences is crucial for ensuring that patients are treated holistically and with fairness in their clinical interactions and in clinical information systems.

Discussion
Although we originally focused on race, gender identity, and sexual orientation, over half of our participants, without probing, shared other types of discrimination. We realized that focusing on racism and LGBTQ+ discrimination was myopic - that understanding the whole person requires acknowledging that there are multiple layers to and many types of discrimination. This brought to our awareness that in healthcare in particular, the impact of many layers and types of discrimination are often left unexamined because they have not yet received the same
attention as medical racism, which can result in gaps in traditional training devoted only to implicit race bias in healthcare. Thus, providers may not recognize or know how to address patients’ intersectional identities. As PT16 commented, “Just in every way that I am different from [my doctor], I think, is a barrier; honestly, the more things we have in common, the more welcomed I have felt so even if that’s just gender sometimes that is even just a huge difference.” This limited understanding doctors have can leave marginalized patients feeling as though they have been taken advantage of and like they are not receiving the best care possible. For example, PT11, a Black queer woman, explained “I complained about knee pain all the time and doctors always said ‘lose weight, lose weight, lose weight’... I got injured on a [theme park] ride and... when the x-rays came back the doctor was like ‘you have arthritis in one of your knees and also your kneecap is misaligned’... I have been dismissed for a decade... maybe I wouldn’t have arthritis in my knee right now had doctors listened to me previously when I expressed pain.” The complexity of PT11’s identity greatly limited the care she received, demonstrating the opportunity for innovative health informatics solutions that can represent patients in their entirety to support doctors in providing equitable, patient-centered care.

Recommendations for the field
How can we, in the biomedical informatics community, move towards patient-centered informatics solutions that empower patients with intersectional identities and encourage holistic treatment? We must prioritize conversations that work to establish what intersectionality means in our field and how we can make intersectionality an expectation rather than an afterthought in our informatics research and practice. Intersectional identities are under-researched as a whole. For example, current demographic categories that inform patient-centered healthcare methods restrict consideration of a patients’ identities to dichotomized dimensions; women versus men, Black versus White, young versus old. These binary categories fail to account for individuals with multiple identities that influence the care they may need. The health informatics community should prioritize ways to reinvent our conventional methods of categorizing individuals in our tools--from Patients Reported Outcomes to Electronic Health Records--and in our research and informatics interventions. Patient-centered informatics solutions can only be successful by considering the entirety of who a person is, as complex as that can be, and not by who they are assumed to be based on their physical characteristics and historic privilege.

Acknowledgements
The study is supported by #1R01LM013301. We want to thank Erin Beneteau, Cezanne Lane, Drishti Vidyarthi, Calvin Apodaca, Steven Rick, Dr. Nadir Weibel, and the rest of the UnBIASED team for their invaluable contribution to the hours of conversations we had surrounding this work and important topic. Most importantly, we would like to thank our participants for sharing their stories.

References
Beyond Absolute Acuity: Changes in Relative Acuity Direct Proactive Rapid Response Teams toward Meaningful Interventions

Joseph Beals, PhD¹, Kathy Belk, BA¹, Sheila Coonan, RN², Jacqueline Laird, RN³

¹PeraHealth Inc., Charlotte, NC; ²Yale New Haven Hospital, Saint Raphael Campus, New Haven, CT; ³Bridgeport Hospital, Bridgeport, CT

Abstract

Early warning technologies increasingly support pro-active clinical rounding on hospitalized patients, but often generate alerts of questionable utility. Simply identifying the most acute patients is unlikely to provide net new information to caregivers. SWAT nurse interventions prompted by the Rothman Index mobile application at Bridgeport Hospital and Yale New Haven Hospital demonstrate that alerts triggered by changes in patient acuity have value that is unrelated to the patient’s overall acuity level. Pro-active rounding teams may be more effectively directed to patients using alerts based on changes in a patient’s acuity without reference to fixed acuity thresholds.

Introduction

As part of the 100,000 Lives Campaign in 2004, the Institute for Healthcare Improvement recommended that hospitals deploy rapid response teams (RRTs), or groups of critical care clinicians trained to act quickly when a patient’s condition declines. By 2010 over half of US hospitals had implemented an RRT.¹ However, the last decade has shown these efforts to have mixed results.²³ Specifically, commonplace failure of the afferent limb – the ability to detect at-risk patients – undermines RRT effectiveness.³ Advances in technology have improved the availability of track-and-trigger early warning score (EWS) systems, however over-alerting from such tools can lead to inefficient use of clinical resources and alert fatigue.⁴ The majority of EWS systems trigger an alert when a pre-defined acuity threshold is crossed.⁵ While threshold-based alerts may correctly identify patients with significant physiological impairment, the condition of such patients is often already known to caregivers and hence the alert provides little incremental value.⁶ Conversely, changes in acuity, even among modestly impaired patients, may not have been appreciated and may be meaningful. This work shows that alerts based on a patient’s relative change in acuity has clinically actionable value across the acuity spectrum.

Methods

SWAT nurses at Bridgeport Hospital (BH) and the Yale New Haven Hospital, Saint Raphael Campus (SRC) used a secure mobile app to receive real-time deterioration alerts for patients on adult, non-ICU inpatient units.⁷ Alerts were based on relative changes, rather than threshold values, of the Rothman Index⁸ (RI) acuity score (PeraHealth, Inc., Charlotte, NC). An RI score of 100 corresponds to no impairment, with decreasing scores implying increasing acuity; RI scores around 40 are typical of patients being considered for ICU level care. Two alerts based on acuity changes were defined: High Alerts for patients whose RI scores dropped by 40% or more over a 12-hour period, and Medium Alerts for patients whose RI scores dropped 30% or more over a 24-hour period. If score changes met both sets of criteria the patients were assigned the High Alert status. Patients with RI scores below 20 were covered by a different alerting rubric so Medium/High warning criteria did not apply. As part of the clinical response protocol, nurses identified actions taken using a structured list, which included over half a dozen resultant interventions spanning a range of urgency from discussion of the patient with the medical team, to the initiation of orders for changes in treatment or additional diagnostics, to calling the full rapid response team to the bedside, calling a code blue, or transferring the patient to a higher level of care. Data from 7,796 alerts from Jan-Oct 2020 were analyzed to determine (a) the acuity level of patients at the time of the alert and (b) how the utility of the alerts (measured by whether or not there was a resulting clinical intervention) varied as a function of patient acuity at the time the alert was triggered.

Results

Owing to alert definitions requiring either a 30% or 40% drop in score over a specified timeframe, RI values of 70 and 60 are the highest scores that can be associated with Medium and High alerts, respectively, of which there were few or no instances. As shown in Figure 1, the preponderance of alerts occurred at lower RI scores (e.g., 96% of High
and 73% of Medium Alerts were associated with RI scores below 45). This is a function of both the baseline (pre-deterioration) acuity of the population as well as the fact that patients with minimal physiologic impairment are generally less prone to significant deterioration.

As shown in Figure 2 the proportion of alerts resulting in an intervention (i.e., alerts with the clearest value) was remarkably consistent across the acuity spectrum, over the RI range of 20 to 60. That is, for a given alert type at a given facility, the likelihood that the alert prompted a change in patient care did not vary with patient acuity.

**Discussion**

Threshold-based acuity alerts implicitly assume that higher acuity patients should be flagged, and that they will benefit from additional clinical attention. In most cases, nurses and physicians are aware of which patients are the sickest, and alerting on these patients provides no new insights. In contrast, deteriorating patients, whether of high or moderate acuity, are the patients whose evolving condition may not be easy to recognize. If it were true that higher acuity patients are more likely to receive interventions on the basis of proactive review, then alerts associated with lower RI scores should show a higher rate of interventions than alerts on less acute patients (those with higher RI scores). In fact, what we see in Figure 2 directly supports the contention that it is the change in a patient’s acuity, regardless of their overall acuity level, that is important to identify.

**Conclusion**

We show that the utility of EWS alerts, as determined by the rate of meaningful subsequent interventions, is independent of overall patient acuity. It is the relative change in any given patient’s acuity, and not alerts focused on absolute acuity thresholds, which can most effectively guide pro-active SWAT or RRT nurse rounding. This is consistent with clinical intuition, and helps explain why alerts using only fixed acuity thresholds are likely to provide an excess of unhelpful alerts on higher acuity patients, while missing important changes in lower acuity patients.

**References**


A Framework to Support Diversity, Equity, and Inclusion within AMIA Through Strengthened Pathways, Support and Leadership

Oliver J. Bear Don’t Walk, IV MS¹; Kevin Wiley, Jr., MPH²; Lois Walters-Threat, DNP, MS, RN-BC³; Rebecca L. Rivera PhD, MPH, CPH²; Martin C. Were, MD MS, FIAHSI, FAMIA⁴; Tiffrani J. Bright, PhD, FACMI⁵

¹Columbia University, New York, NY; ²Indiana University, Indianapolis, IN; ³American Nurses Credentialing Center, Silver Spring, MD; ⁴Vanderbilt University Medical Center, Nashville, TN; ⁵IBM Watson Health, Cambridge, MA

Introduction

In June 2020, the AMIA Board of Directors established the Diversity, Equity, and Inclusion (DEI) Task Force (now Committee) to advise AMIA on specific, actionable steps addressing DEI matters for strategic planning. An immediate charge of the DEI Task Force was to recommend a framework to support workforce diversity, exposure to biomedical informatics, awareness of AMIA, and improved recruitment and retention of students and faculty. Objectives were to strengthen pathways into biomedical informatics and address the unmet needs of AMIA members who are part of marginalized or historically excluded groups. This was in recognition that achieving a diverse and inclusive organization and environment would enable better and bolder decision-making while improving outcomes for the organization. Philips states, “Diversity jolts us into cognitive action in ways that homogeneity simply does not⁴.” Successful implementation of the proposed framework would ultimately benefit all members through collaboration.

Methods

To develop the framework, DEI sub-committee members conducted desk-reviews of available literature. Additionally, a brief scan of comparable professional association reports and programs for DEI initiatives was conducted for comparison. Scanned professional associations were of similar size to AMIA, focused on STEM and/or medicine, and had task forces or committees drafting actionable DEI strategies. Iterative sessions of brainstorming, editing and organizing activities were conducted over nine months to identify potential activities. A consensus-based approach was then used to prioritize identified activities needed to meet the specified objectives. Activities were represented using logic models⁵, which ensured that activities, resources, outputs and outcomes were comprehensively considered. Outputs are direct, tangible results of the activities. Outcomes are the desired results of actions. Logic models flow from activities to outcomes. Activities represent concrete steps AMIA can take to improve DEI.

Results

Comparable organizations have developed diversity initiatives to improve overall representation of marginalized or historically excluded members. Many DEI efforts were recently initiated and usually emphasized one point of the DEI triad. A common theme among scanned organizations was a focus on improving pathways into their respective fields by targeting students and trainees. These efforts directed organizational resources to develop strategies and committees to organize and mobilize members to improve DEI within each discipline. In all, the DEI initiatives of five other professional organizations were compared to the proposed framework.

Overall, there are 23 activities spread across three logic models: (a) Recruitment & Retention, (b) Supporting workforce diversity, and (c) Exposure to Informatics. The Recruitment & Retention activities contained 12 recommendations; Supporting workforce diversity contained eight recommendations; and Exposure to Informatics contained three recommendations. Recruitment & Retention activities pertain to building mutually beneficial partnerships with and pathways into AMIA for institutions and organizations that serve marginalized and historically excluded communities. It also brings into focus intentional leadership and mentorship for marginalized and historically excluded members. Supporting Workforce Diversity domain has an initial aim of building transparency and knowledge about the diversity climate within AMIA. Exposure to Informatics has one main goal to fortify existing and creating new outreach programs with broader scopes and partnerships. Here, two action items are presented for each of the three logic models outlined in Tables 1, 2, and 3.

Conclusion

The logic-model-based framework developed here provides a roadmap to guide AMIA in actualizing its goal of promoting diversity, equity, and inclusion of its membership. Making AMIA a welcoming home for all would take thoughtful and targeted actions to correct barriers that continue to keep many communities from being represented in informatics more generally. Some activities will have outcomes that can be realized within a year, and others may
take years for measurable results. A commitment to DEI is an ongoing process of unlearning these systems of oppression and then learning alternatives that promote the wellbeing of all. The proposed recommendations align with our field’s commitment to the ethics outlined in the Belmont Report\(^1\). To authentically realize these principles, they need to extend beyond the lab setting. How can respect for persons, beneficence, and justice be truly realized in our research without first addressing systems of oppression which violate these principles within our own community? Even if one is only concerned with ethics in research, a commitment to these principles necessitates a diverse, equitable, and inclusive community as a foundation.

**Table 1-Logic Model for Retention and Recruitment**

<table>
<thead>
<tr>
<th>Activities</th>
<th>Outputs</th>
<th>Outcomes</th>
<th>Metrics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Targeted outreach and communications to institutions that serve marginalized and historically excluded communities and BMI departments with DEI efforts.</td>
<td>Continual communication with these institutions on upcoming conferences, research, calls for research and opportunities</td>
<td>Increased participation in AMIA from these institutions</td>
<td>Number of conference attendees, and research submissions and acceptances from these institutions and DEI programs</td>
</tr>
<tr>
<td>Leadership and mentorship program with a focus on marginalized and historically excluded members</td>
<td>Mentorship and leadership program developed</td>
<td>Increased number of AMIA members from marginalized and historically excluded communities in leadership positions and paired with mentors through AMIA</td>
<td>Number of potential mentors from marginalized and historically excluded communities</td>
</tr>
</tbody>
</table>

**Table 2-Logic Model Supporting Workforce Diversity**

<table>
<thead>
<tr>
<th>Activities</th>
<th>Outputs</th>
<th>Outcomes</th>
<th>Metrics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collect and share more precise &amp; transparent data on the diversity of AMIA membership, in addition to the metrics proposed here.</td>
<td>Data on diversity in AMIA</td>
<td>Targeted DEI actions by AMIA through knowledge of how well different groups are represented</td>
<td>Annual reports on the diversity of membership papers, presentations and posters and overall changes from previous years</td>
</tr>
<tr>
<td>Listening sessions to understand membership turnover in marginalized and historically excluded communities</td>
<td>Data on member turnover in AMIA, especially within marginalized and historically excluded communities</td>
<td>Targeted DEI actions by AMIA through knowledge of members’ reasons for joining/staying/leaving</td>
<td>The presence of listening sessions conducted every 5 years</td>
</tr>
</tbody>
</table>

**Table 3-Logic Model for Exposure to Informatics**

<table>
<thead>
<tr>
<th>Activities</th>
<th>Outputs</th>
<th>Outcomes</th>
<th>Metrics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Design mechanisms for current members to serve as liaisons to conferences that serve marginalized and historically excluded communities in STEM.</td>
<td>Liaison funding and information for individuals</td>
<td>AMIA members with financial and logistical support to serve as liaisons at conferences which serve marginalized and historically excluded communities</td>
<td>Number of liaisons supported</td>
</tr>
<tr>
<td>Develop AMIA’s role as a communicator between academia &amp; industry and the broader community</td>
<td>Methods, programs and partnerships to disperse research knowledge about health equity</td>
<td>Relationships with marginalized and historically excluded communities focused around healthcare topics</td>
<td>Number of partnerships with communities to disperse knowledge</td>
</tr>
</tbody>
</table>

**Acknowledgments.** The authors thank the DEI Task Force members, AMIA staff and especially the Building a Diverse AMIA subcommittee for their insights, help and support.

**Learning Objectives**

At completion of this activity, participants will be able to (1) Understand how the AMIA Diversity, Equity, and Inclusion Task Force addressed the DEI Charge given by the AMIA Board to create a framework that strengthens pathways into informatics and AMIA and address the unmet needs of members from marginalized or historically excluded communities.

**References**


Forecasting Inpatient Bed Demand to Respond to the COVID-19 Pandemic

Biplab Sudhin Bhattacharya PhD1, Debdipto Misra MS1, Eric Reich MSHI1, Jason Puckey BS1, David Vawdrey PhD1
1Geisinger Steele Institute for Health Innovation, Danville, PA, USA

Introduction – The COVID-19 pandemic caused extreme swings in hospital occupancy across the United States. Expecting a surge in COVID-19 cases in the spring of 2020, our institution canceled elective surgical procedures, made surge beds available, and redeployed staff. As the first wave of COVID-19 cases receded, a large backlog of non-urgent surgical cases needed to be scheduled—a proposition that threatened to overwhelm inpatient bed capacity. As we faced a second COVID-19 wave in the fall and winter that was 3-4x more severe than what we experienced in the spring, managing bed availability by titrating scheduling of surgical cases became even more important.

To help with short- and long-term management of inpatient capacity and operating room scheduling, we developed a time series Seasonal Autoregressive Integrated Moving Average (SARIMA) model. At the early stages of the pandemic, predicting COVID-19 specific admissions was a difficult task1. We used this model to produce 7-day occupancy forecasts for all medical, emergent surgical, and non-emergent surgical inpatient census for eight hospitals. The occupancy forecasts were distributed daily to clinical and operations leaders to aid decision-making in areas of discharge planning, staffing and surgical scheduling.

Methods – Each day for eight hospitals, the forecast models were retrained using the most recent inpatient occupancy counts. Occupancy was calculated as the number of medical, emergent surgical, or non-emergent surgical patients occupying inpatient beds, plus the number of patients that were being held in alternate locations such as the emergency department, post-anesthesia care unit, or cardiac recovery suite. The occupancy data were obtained from the bed management system in the institution’s electronic health record (Epic Corp., Verona WI).

There are various methodologies that can be used to predict occupancy2. Before the pandemic the analytics team at Geisinger had been developing a LSTM (Long Short-Term Memory) neural network approach which was selected after being tested against other time series models (ARIMA, exponential smoothing and SVR). The pandemic changed the statistical characteristics of hospital occupancy and the LSTM approach started performing poorly (Table 1). SARIMA models are univariate models, that are more responsive to unexpected fluctuations in the data, and are easier to implement and deploy.

We implemented three SARIMA models to predict occupancy 1-7 days in the future for: 1) medical, 2) emergent surgical, and 3) non-emergent (elective) surgical patients. The seasonal component of the model accounted for weekly patterns in bed occupancy (e.g., more inpatients on weekdays vs. weekends). Figure 1 shows an example of high variation in occupancy at one of our hospitals at the outset of the pandemic (beginning mid-March 2020). Such variation is indicative of non-stationarity in the input data, caused here by mass cancelation of surgical cases, declining medical inpatient census for non-COVID-19 patients, and unpredictable volumes of COVID-19 admissions. To account for this non-stationarity, an auto-parametrizing technique3 was implemented, which selected the optimal parameters for the SARIMA models each time they were run. Model performance was measured using mean absolute error (MAE) and mean absolute percentage error (MAPE).

It is vital that predictive models produce actionable insights4. The framework we used for deployment is shown in Figure 2. The models ran on an on-premise Hadoop Cluster which houses the institution’s enterprise data warehouse. An automated daily process tabulated medical, emergent surgical and non-emergent surgical occupancies that were used as model inputs. An automated report was generated (Figure 3) and sent by email to several hundred stakeholders, including executives and clinical and operations leaders. Additionally, the forecast data were included within other dashboards across the health system.

Results – Table 2 shows the SARIMA model performance for total bed occupancy from September 01, 2020 to October 30, 2020 for Geisinger Medical Center, an academic medical center with 367 adult beds. The MAE (actual bed occupancy vs. predicted) was between 13.7 and 16.8 beds for 1-7 days in the future. Overall MAPE for the evaluation period was 4.7%.

Discussion – Simple ARIMA models were among the first tools to predict bed occupancy5. Using a seasonal variant to account for weekly fluctuations along with the auto-parametrizing feature based on the stepwise algorithm1 enabled the model to adjust itself to the wide fluctuations in occupancy we observed during the COVID-19 pandemic. This model adapted to these fluctuations much better compared to a more sophisticated neural network model.

There are significant costs associated with overstaffing, while suboptimal patient care and dissatisfaction occur because of understaffing. The daily distribution of occupancy forecasts generated by our models allowed clinical and operations leaders to make data-driven decisions in planning for resources, staffing, discharges, and surgical scheduling. The availability of accurate
Forecasts of 7-day census became a key tool for Geisinger in battling the COVID-19 pandemic, and we expect the models to be equally valuable in the future to manage bed capacity in an increasingly challenging healthcare delivery environment.

**Figures and Tables**

**Figure 1.** Inpatient occupancy at Geisinger Medical Center during the first wave of the COVID-19 pandemic.

![Inpatient occupancy graph](image1.png)

**Figure 2.** Implementation framework for the SARIMA occupancy forecasting models.

![Implementation framework](image2.png)

**Figure 3.** Snapshot of daily occupancy forecast sent to key stakeholders at 367-bed Geisinger Medical Center.

![Snapshot of forecast](image3.png)

**Table 1.** Deterioration of performance for the LSTM neural network model before and during the COVID-19 surge. Results for before COVID-19 are for the timeframe 8/15/2019 – 1/31/2020, and for during COVID-19 surge are for June 2020.

<table>
<thead>
<tr>
<th>Time frame</th>
<th>MAE (beds)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 day ahead</td>
</tr>
<tr>
<td>During COVID-19 surge</td>
<td>25.36</td>
</tr>
</tbody>
</table>

**Table 2.** Predictive MAE and MAPE performance of the 1- to 7-day ahead model for 367-bed Geisinger Medical Center.

<table>
<thead>
<tr>
<th>Prediction</th>
<th>MAE (beds)</th>
<th>MAPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Total Census)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1-day ahead</td>
<td>13.73</td>
<td>3.90%</td>
</tr>
<tr>
<td>2-day ahead</td>
<td>14.4</td>
<td>4.20%</td>
</tr>
<tr>
<td>3-day ahead</td>
<td>16.2</td>
<td>4.90%</td>
</tr>
<tr>
<td>4-day ahead</td>
<td>16.3</td>
<td>5.00%</td>
</tr>
<tr>
<td>5-day ahead</td>
<td>16.78</td>
<td>5.20%</td>
</tr>
<tr>
<td>6-day ahead</td>
<td>15.95</td>
<td>4.80%</td>
</tr>
<tr>
<td>7-day ahead</td>
<td>16.23</td>
<td>4.70%</td>
</tr>
<tr>
<td>Aggregate 7-day performance</td>
<td>15.65</td>
<td>4.70%</td>
</tr>
</tbody>
</table>

**References**

LVHNet: Detecting Cardiac Structural Abnormalities with Chest X-Rays

Shreyas Bhave, MA1 Pierre Elias, MD2, Victor Rodriguez, MPhil1, Timothy Poterucha, MD3, Simi Mutasa, MD2, Jay Leb, MD3, Nir Uriel, MD2, Adler J Perotte, MD MA1
1Columbia University, Department of Biomedical Informatics, New York, NY, USA; 2Columbia University Irving Medical Center, New York, NY, USA

INTRODUCTION. Early identification of structural changes to the heart such as left ventricular hypertrophy (LVH) and dilated cardiomyopathy (DCM) is critical to improving outcomes in heart failure. Initial signs and symptoms of early stage heart failure can be nonspecific, and there will often be a delay in diagnostic echocardiography while fast, inexpensive chest X-rays (CXRs) are much more frequently performed. In the same way breast cancer screening mammograms evolved from chest radiographs, there is promise that CXRs can form a basis for early heart failure detection. Deep learning models utilizing CXRs have lacked both variety and quality in cardiac labeling. Prior models have primarily focused on detecting cardiomegaly, a catch-all term for an abnormally enlarged heart and often the only statement radiologists will make about the heart. Despite this, the radiologic comment of cardiomegaly is known to be poorly predictive of cardiac disease and does not trigger meaningful clinical action. We sought to replace cardiomegaly with more accurate and clinically actionable labels of structural abnormality derived from gold standard echocardiograms. These structural abnormalities lead to enlargement of the heart and are the key changes that occur before the development of overt heart failure. The detection of structural abnormalities consistent with LVH and DCM phenotypes could lead to earlier diagnosis of cardiac pathology, improving the management of heart failure.

DATASET AND LABELS. We created a dataset of 88,769 echocardiograms and 49,620 posteroanterior (PA) CXRs performed within 12 months of one another on 16,579 unique patients from Columbia University Irving Medical Center. CXR images and tabular data from echocardiograms were extracted, including LV end-diastolic dimension (LVEDD), end-diastolic interventricular septal thickness (IVS), end-diastolic posterior wall thickness (PWT). We constructed three gold standard labels using the echocardiogram data corresponding to each CXR: LVH, Severe LVH (SLVH) and DCM. The presence of positive labels was: 13,156 for LVH (26.3%), 4,326 for SLVH (8.6%) and 1,583 for DCM (3.2%). From this data, we constructed a train (N=38943, P=13059), evaluation (N=4767, P=1632), and test (N=4648, P=1633) split.

MODEL ARCHITECTURE, PREPROCESSING AND FEATURE MAPS. CXR images were cropped to a 1:1 aspect ratio, downsampled using bicubic interpolation, and normalized using contrast limited adaptive histogram equalization and minimax normalization. In order to improve the generalization performance of the model, we augmented our dataset by adding Gaussian noise (mean=0, var=0.05). All experiments were conducted using DenseNet-121. We devised two separate tasks: (1) multi-task binary label prediction on all three labels and (2) prediction on a single composite label which represents presence of LVH, SLVH or DCM. We used gradient-weighted class activation maps (Grad-CAM) to produce heatmaps visualizing areas of the images which were most responsible for driving prediction.

![AUC-ROC and AUPRC Curves for both multi-task and composite models](image)

*Figure 1 AUC-ROC and AUPRC Curves for both multi-task and composite models*
**COMPARISON TO RADIOLOGISTS.** We compared LVHNet's performance on detection of cardiac structural abnormality to radiologists commenting on presence of cardiomegaly via radiology reports with the label being extracted on the entire test set via CheXbert. We manually reviewed a random sample of the test set radiology reports to confirm accuracy of the labeler and remove ambiguous cases. Cardiac structural abnormality was defined as the presence of LVH, SLVH or DCM as identified via echocardiography. In discussion with chest radiologists, consensus was achieved that the fairest comparison to the radiologist label of cardiomegaly would be the composite label.

**RESULTS: MODEL PERFORMANCE.** We report the ROC and PR curves for the model in the multi-task and composite label setting (Figure 1). The multi-task model exhibits the best AUC-ROC on DCM (0.833) and strong performance on LVH (0.778) and SLVH (0.795) as well. The model performs best overall on the LVH label as the AUPRC is significantly better (0.554) in comparison to the other labels. The model maintains a relatively good AUC-ROC (0.774) and AUPRC (0.603) on the composite task.

**RESULTS: FINDINGS FROM GRAD-CAM.** The multi-task model Grad-CAM heatmaps demonstrate that for true positives, the model consistently focuses on structures of the heart across all structural abnormalities. Meanwhile, the true negative heatmaps highlight areas in the thoracic/abdominal cavities with no propensity for focusing on the heart.

![Figure 2 Grad-CAM heatmaps for targeted cardiac diseases. Blue borders: true positives, Orange borders: true negatives](image)

**RESULTS: PERFORMANCE COMPARED TO RADIOLOGIST CARDIOMEGALY READS.** Lastly, we compared the model’s ability to accurately detect cardiac structural abnormalities versus the official diagnostic read from a radiologist (Table 1). LVHNet outperforms radiologists in detecting structural abnormality on accuracy and F1 score. The accuracy of radiologist reads of cardiomegaly with CXRs was 70.5% while echocardiography has been consistently shown to have accuracy in excess of 97% at diagnosing structural abnormality. Our model is able to achieve closer to ground truth accuracy by training on these higher quality labels. These results demonstrate a significant advancement in the ability to detect cardiac structural abnormality from chest X-rays and the benefit of leveraging labels which are more specific indicators of cardiac pathology than cardiomegaly. Since state-of-the-art deep learning models using CXR data all depend on the cardiomegaly label, LVHNet represents an important step towards detecting actionable cardiac structural abnormality from CXRs.

<table>
<thead>
<tr>
<th>Label</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Radiologist Cardiomegaly</td>
<td>.705 (.689, .719)</td>
<td>.515 (.487, .545)</td>
<td>.568 (.534, .595)</td>
<td>.541 (.512, .567)</td>
</tr>
<tr>
<td>Model Composite Prediction</td>
<td>.742 (.726, .758)</td>
<td>.566 (.538, .595)</td>
<td>.664 (.632, .691)</td>
<td>.611 (.586, .633)</td>
</tr>
</tbody>
</table>

**References**

Multi-Channel LSTM for Modeling Irregularly Sampled Time Series

Shreyas Bhave, MA1, Adler J Perotte MD MA1

1Columbia University, Department of Biomedical Informatics, New York, NY, USA

INTRODUCTION. Multivariate irregularly sampled time series data are ubiquitous in many data modalities across healthcare including principally Electronic Health Records (EHR) data. They are defined in the context where a dataset contains a set of time series where there are D distinct features, the time intervals between successive observations are irregular and all D features are not always present at each successive time point. Deep learning approaches that deal with modeling time series data such as RNNs and LSTMs have achieved state of the art results on many prediction tasks. However, these models are primarily used for regularly sampled data and are based on regular sampling assumptions. Furthermore, they are known to struggle in modeling signals that occur at different timescales simultaneously. In particular, they struggle to capture long-term dependencies in the data which may inform prediction. EHR data consists of data streams observed at very different rates along with hospital visits which occur at irregular intervals. Moreover, important downstream prediction tasks of interest such as in-hospital mortality, decompensation, and length of stay rely on effectively modeling the full history of a patient’s visit. In order to address these issues, we propose the Multi-Channel LSTM which consists of D separate LSTM channels, one for each feature, which are tasked with learning the irregular dynamics of each signal separately and combining all the signals to produce a prediction. We tackle the challenge of irregular sampling within each signal by incorporating information about the time gaps between data points directly. We show that this model outperforms baselines on a decompensation prediction task using MIMIC-III data.

PAST WORK. There are broadly two approaches that past work has taken: (1) Discretize the irregularly sampled sequence into equal bins and develop an interpolation model for data that is missing prior to using a standard RNN1,2 (2) Incorporate the time gaps between subsequent data points directly in the RNN by modeling the dynamics of the hidden state3,4. We propose an approach which is closer to (2) with the main distinction being that we model each feature separately, incorporating time differences between subsequent observations of the same feature.

METHODS: Multi-Channel LSTM (MC-LSTM). An unrolled representation of this model is shown in Figure 1. In this figure, there are three separate channels (one for each feature) which have their own LSTM. Each channel is trained using both the data observed at each time as well as the time difference between the current and previous timestep. Outputs can be predicted at any time. For prediction at any given time, we extract the hidden states of all channels which encode a representation of the entire history of the irregular time series and use a feed forward neural network using the concatenated hidden states to combine the signals. Such a model has a number of theoretical advantages: (1) It is amenable to modeling features which occur at vastly different timescales, mitigating the problem of vanishing gradients that is typically a challenge for such time series models. (2) It does not require building an imputation model which can be challenging especially when data are missing not at random.

METHODS: Hierarchical Multi-Channel LSTM (H-MC-LSTM). The principal drawback of the MC-LSTM is that it does not directly model dependencies across the features overtime as a model which leverages a multivariate vector representation and a single LSTM does. In order to remedy this issue, we introduce a hierarchical version of MC-LSTM which collects the hidden states across all channels and uses them as input to a regularly-spaced LSTM. This regularly-spaced LSTM contains information about the relationships between features and the overall dynamics of the hidden states across all channels. An unrolled representation of this model is shown in Figure 2. In this figure, there are three separate channels and one higher-level regularly spaced channel. This regularly-spaced channel incorporates all of the latest hidden states and times since they were observed for all the channels at a user-specified regular interval. Subsequently, this regularly-spaced channel’s hidden state is used along with the channel-specific
hidden states for prediction at any given time.

**DATA AND EXPERIMENTS.** In order to evaluate the model, we make use of the mimic benchmarks\(^5\) task of decompensation. The task is to predict whether a patient’s health will rapidly deteriorate in the next 24 hours. In particular, the task is a binary prediction task at each timestep from admission to the end of the ICU stay, where a positive label occurs in all timesteps 24 hours prior to death. Here, death is used as a proxy for decompensation as it is a relatively easy outcome to accurately label in the MIMIC-III\(^6\) database. We chose this task because it is widely used in the community, is of particular interest for early warning scoring and is supervised at each timestep. It also consists of a set of 17 physiological features which are sampled at varying timescales (e.g. heart rate, glucose, pH). The baselines we compared against for this task were a channel-wise LSTM, a discretization strategy used in the original mimic benchmarks paper (their best model) and our own vanilla LSTM model which is discretized and imputed using fill forward imputation.

**RESULTS AND DISCUSSION.** The results in Table 1 show that the MC-LSTM and H-MC-LSTM both outperform the two baselines on the AUC-PR metric. The channel-wise LSTM has the same model capacity as MC-LSTM but uses a discretization strategy. This shows that the performance improvement is not a result of a difference in the number of model parameters. The MC-LSTM is better able to capture the signal from each individual time series and avoids introduction of noise due to discretization and imputation techniques. Thus, the MC-LSTM shows great promise for modeling irregularly sampled data. In the future, we hope to further test this model in other risk prediction settings where both short and long term signals may be important.

**Table 1 Results on two baselines versus MC-LSTM and H-MC-LSTM (bolded results significantly outperform baselines)**

<table>
<thead>
<tr>
<th></th>
<th>AUC-ROC</th>
<th>AUC-PR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vanilla LSTM Imputed</td>
<td>0.901 (0.898, 0.905)</td>
<td>0.345 (0.335, 0.355)</td>
</tr>
<tr>
<td>Channel-wise LSTM</td>
<td>0.911 (0.908, 0.913)</td>
<td>0.344 (0.334, 0.354)</td>
</tr>
<tr>
<td>(Harutyunyan et. al.)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MC-LSTM</td>
<td>0.906 (0.903, 0.909)</td>
<td>0.362 (0.352, 0.372)</td>
</tr>
<tr>
<td>H-MC-LSTM</td>
<td>0.911 (0.907, 0.914)</td>
<td>0.375 (0.369, 0.384)</td>
</tr>
</tbody>
</table>

**References**

Latinx Acculturation Scales: Scoping Review and Novel Scale Selection Tool

Anweysha Bhowmik*, MS1, Logan Cameron*, BS1, Julia Ivanova, MA1, Stanley Haynes, MAS1, Ariana Cano, MAS1, George Runger, PhD1, Anita C. Murcko, MD1
1Arizona State University, College of Health Solutions, Scottsdale, AZ
*These authors contributed equally

Introduction: “Acculturation” has been defined as the process by which cultural groups adopt the customs and behaviors of a culture²³. The concept of acculturation is dynamic, moving from linear constructs like assimilation, into complex multivariate concepts such as biculturalism, multiculturalism, and cultural pluralism. Though numerous scales in multiple domains now exist, few studies quantify and categorize acculturation scale usage, so researchers seeking scales specific to their discipline or population are often limited to the validation study or a qualitative analysis. This study applies the National Institute on Minority Health and Health Disparities (NIMHD), a framework¹ that facilitates analysis of minority health and health disparities research portfolios to assess progress, gaps, and opportunities. Visualizations are based on the Sankey diagram popularized by Tufte⁴, correlating proportion to the diagrammed flows of information.

Method: PubMed, Scopus, and PsycInfo were searched using an optimal Boolean search string for articles published in the US 2015-2020. Books, editorials, and literature reviews were excluded. After eligibility screening, 32 articles remained. Multidisciplinary researchers categorized articles and acculturation scales using the NIMHD framework¹. Zotero was used for data storage, organization, and data preprocessing. Using a contingency table of domains, population size and acculturation scales, a Sankey-based tool for selection of acculturation scales was created using Tableau⁵.

Results: From the scoping review we identified and categorized 17 validated Latinx acculturation scales. We note that such scales must be interpreted in the context of geographic influences, generation of immigration, and other socioeconomic factors⁵. This standardized data on validated acculturation scales (click here⁷ for table) was transformed via the Tableau Sankey diagram software, as illustrated in Figure 2. A novel decision support tool, Latinx Acculturation Scale Selection Tool, was created based on Sankey Tableau software that permits users to input population and research domains, apply filters for influence and health outcome domains and identify patterns in their area of interest, thereby optimizing Latinx acculturation scale(s) selection.

Conclusion: Using the NIMHD framework, we analyzed the past 5 years of Latinx acculturation scale literature to create the novel, interactive, Tableau-based Latinx Acculturation Scale Selection Tool with filters for influence and health outcome domains. Researchers and health plans can use the Tool to retrieve acculturation scale use patterns in their area of interest, thereby informing tool selection based on user-defined parameters.
Figure 1. Flowchart of literature search and screening process.

Figure 2: A Sankey diagram for classification of Latinx acculturation scales based on NIMHD framework.

References:
7. Data Table for Sankey Diagram: https://docs.google.com/document/d/1qKn0NkrNTdNexL05JMATxJ-wzWayqwAc78XIDTkF/edit?usp=sharing
Augmenting Barbershop Initiative Health Promotion with Wearables in African-American Populations
Lisa Anne Bove, DNP, RN – BC, Stephanie Turrise PhD, RN-BC, APRN, CNE, CHFN-K, FAHA and Rachel Carroll, PhD University of North Carolina Wilmington, Wilmington NC, United States

Introduction
According to the 2019 American Heart Association Heart Disease and Stroke Statistics, non-Hispanic African-American men have higher rates of hypertension (HTN) and the highest rate of HTN-related deaths of any racial, ethnic, or gender group in the United States. African-Americans develop cardiovascular disease (CVD) at a younger age than other ethnic groups, and because they remain undiagnosed longer, they experience increased morbidity (heart disease, kidney disease, eye disease) and mortality. Reducing systolic blood pressure (SBP) significantly reduces the risk of CVD and all-cause mortality. Exercise is significantly associated with blood pressure (BP) reduction regardless of the individual’s initial BP level, gender, physical activity level, type of exercise performed, or exercise training program.

A number of studies have demonstrated health improvements due to monitoring health through activity trackers such as Fitbits®. Increased physical activity reduces BP and activity trackers have been shown to increase activity. Thus, wearing an activity tracker may reduce BP. While 19% of Americans reported using an activity tracker such as a Fitbit® or a smartwatch, low income, African-Americans are least likely to use them.

Health promotion interventions have been implemented in barbershops to help reduce risk factors and promote health. Barbershops are often community centers serving diverse populations. Barbers and stylists are willing participants in health promotion activities and help enroll their customers in research studies. Victoria and colleagues found significantly higher rates of HTN control in the experimental group than the control group at the 10-month follow-up of their barbershop intervention. Likewise, Hess and colleagues found that when barbers offered encouragement and told their success stories, their clients who received the increased exposure also demonstrated higher rates of HTN control. While this form of health outreach is becoming more common, there are no studies reporting on combining health promotion with the use of wearable activity trackers as part of a barbershop initiative.

The purpose of this study was to determine feasibility of an activity tracker intervention to measure heart rate, steps, and sleep in African-Americans to determine 1) if the use of these activity trackers impacted health behaviors and BP, 2) the relationship among wearing patterns, steps, heart rate, sleep, and BP, and 3) any technical or wearability issues with the Fitbits®.

Methods
African-American barbers and stylists who are part of a regional medical center’s Barbershop Initiative were recruited from three shops. Participants signed consent, completed an initial survey, and were provided assistance setting up their Fitbit Inspire HR®. Participants wore the Fitbit® for a month collecting heart rate, steps, and sleep data. Weekly BPs, wearing patterns and health questions were collected. Upon study completion participants completed another survey describing their wearing patterns and perceived impact to health. Descriptive and inferential statistics including multiple linear regression and t-tests using R® were conducted.

Results
Out of 14 participants, the majority were males (71%) with known HTN (64%), although only 29% took antihypertensive medications. Age ranged from 25-75 years (M=46, SD = 17) and body mass index (BMI) ranged from 19–46 (M=30, SD =8.5). Preliminary results suggest that over the 30-day period, users averaged 7876 steps a day with an average resting heart rate of 72 beats per minute. The average change in systolic BP ranged from -22 to 34 (M=2.46) and diastolic BP ranged from -11 to 13 (M=3.6) where negative values indicate worse blood pressure compared to baseline. All but one participant reported wearing the device all or most of the time. All participants wore the device to sleep, though not every night. Nearly 30% (N=4) wore a tracker before, but admitted they did not monitor the results.
Participants also reported they were exercising more (N=7, 50%), getting more sleep (N=5, 36%), and felt they had better blood pressures (N=10, 71%). Anecdotally, participants reported that they ‘paid attention’ to the automated reminders and alerts and some reported increasing exercise based on the reminders.

Discussion

While there were no statically significant changes in BP, participants reported more activity and a focus on improving their risk factors for CV disease. Researchers were able to access information from the trackers while the participants were part of the study. While HR and sleep data was captured, it was difficult to correlate without hourly data extracts. Since participant's smartphone minutes were not unlimited, they often turned off Bluetooth, and so needed reminders to synchronize their Fitbits with the application on their phone each time we met with them. Adding an application program interface (API) to extract the data will improve quality of data for statistical analysis for meaningful interpretation about the relationship between HR and sleep.

References

Developing an Academic-Industry Internship to Train Next-Generation Biomedical Informaticians

Tiffani J. Bright, PhD, FACMI1, Allison B. McCoy, PhD, FAMIA2,
Dilhan Weeraratne, PhD1, Kim M. Unertl, PhD, FACMI2
1IBM Watson Health, Cambridge, MA; 2Vanderbilt University Medical Center, Nashville, TN

Introduction

Academic-industry partnerships are vital for enabling scientific discovery and for workforce development. One of the most beneficial aspects of these partnerships is the opportunity for joint industry-academic internships. Internships allow students to apply theoretical knowledge, gain insight, and learn key job skills. Particularly for undergraduate students with an interest in the science, technology, engineering, and math (STEM) fields, these experiences can help to inform career decisions and develop research skills. Today, students pursuing biomedical informatics and data science internships are faced with a wealth of opportunities across career sectors. However, siloed experimental learning experiences can leave students inadequately prepared to learn how to transfer foundational knowledge and skills across domains. Formal academic-industry internships that cross-train students within a discipline can expand students’ career preparation, benefitting both the individual and the scientific ecosystem. Despite these benefits, formal academic-industry biomedical informatics internships are rare. We describe an academic-industry internship between Vanderbilt Medical Center University (VUMC) and IBM and early lessons from its implementation.

Methods

Student Selection. The internship was created as part of the established Vanderbilt Biomedical Informatics Summer Program (VBISP), with additional positions funded through the IBM-VUMC academic research collaboration to advance artificial intelligence (AI) in healthcare. The VBISP internship was open to undergraduate students from across the USA; underrepresented minorities were encouraged to apply. Candidate applications included personal information, transcripts, reference letters, and a resume. Due to COVID-19, VBISP was significantly scaled back for summer 2020, leading to a need for a novel approach to continue the joint IBM-VUMC internships. Rather than hosting IBM-VUMC funded students in the summer, the internship resumed virtually during the 2020-2021 academic year. VUMC reviewed applications from individuals who had applied for the summer as well as additional applications to identify potential candidates for this academic-year program. The VUMC principal investigator (PI) screened applications according to the VBISP process, identified candidates with related skills for the IBM-VUMC projects, and sent applications to the IBM PI for review. When possible, both PIs interviewed and discussed the candidates to make the intern selections. PIs developed projects to align with intern skillset and interest; when IBM solutions were involved, the IBM scientist took the lead, and VUMC faculty were matched based on expertise and research interests.

Program Components. Internship objectives were to: 1) expand knowledge of biomedical informatics and research skills in academia and industry, 2) provide experiential learning through direct research and industry engagement, 3) establish next steps in the field, and 4) provide opportunities for professional development. The program lasted between 7-10 weeks. Interns gained insight into academic career paths by participating in research and methodological training workshops, twice-weekly seminars, one-on-one mentorship in student research projects, and intern group events. Exposure to life as an industry scientist, which was unique to the IBM-VUMC interns, included participation in the IBM Center for AI, Research, and Evaluation (an industry research team), bi-weekly research meetings, a visit to IBM Watson Health headquarters in Cambridge, MA, and access to IBM scientists. In lieu of the IBM visit due to pandemic travel restrictions, the PIs increased opportunities to receive broader exposure to multi-disciplinary industry sectors by including additional IBM research seminars.

Governance. The joint internship was overseen by one PI from IBM (TJB) and one from VUMC (KMU). The VUMC PI provided oversight for the VBISP, serving as the main contact for the IBM-VUMC Internship. The PIs were assisted by a VUMC staff member who aided with tasks such as program announcement, student recruitment, logistical arrangements, and program evaluation. Interns were guided throughout the research process and co-mentored by a
VUMC faculty mentor and IBM scientists via weekly meetings and ad-hoc meetings, as necessary. Lastly, debriefing sessions were held at the conclusion of the internship.

Results

Since 2019, the program has hosted six undergraduate interns: two in 2019 and four in 2020. Across the cohorts, the majority have been female (67%), Asian (67%), rising seniors (50%), and Vanderbilt University students (67%). One intern was a first-generation college student. The 2019 cohort interns collaborated on a systematic review evaluating the impacts of implemented AI-enabled clinical decision support; both gave an oral presentation at the VBISP Presentations Seminar, presented posters at the Vanderbilt Summer Science Academy Annual Research Symposium, and had poster abstracts accepted at the AMIA 2020 Informatics Summit. The 2020 cohort projects focused on: 1) disparities in genomics testing for lung cancer; 2) sociodemographic disparities in access to and use of telemedicine before and during the COVID-19 pandemic; 3) user interpretation and processing of ambiguous text in clinical notes; and 4) machine learning to assist with COVID-19 resource allocation. At conclusion, each intern gave an oral presentation for IBM and VUMC and submitted a conference abstract. Interns developed research skills, obtained work experience, and learned how scientific research informs product development and evaluation.

We observed several key lessons. First, administrative processes and contractual details required considerable time and needed to be addressed early in the process. Discussions were required to identify roles and responsibilities; align university, medical center, and industry human resources and onboarding policies; determine policies regarding data and intellectual property provisions; and negotiate indemnification and liability. The existing infrastructure for VBISP assisted with navigating these process-related program aspects. Once processes were established in 2019, the subsequent year only required minor modifications, even with the changed format and timing. Second, coordination across academia and industry was necessary to facilitate successful program execution. This was achieved by defining and adhering to responsibilities, regular communication, use of collaborative tools for project management and data sharing, and weekly meetings. Each mentor-mentee team also completed and submitted a project agreement and project plan, providing additional structure and alignment. Lastly, ample financial and personnel resources were essential. IBM sponsored the internship, and VUMC provided support for VUMC personnel for the additional effort required for execution of the joint internship. Small, but key levels of effort should be allocated for project management administration. Budgetary flexibility is also necessary to accommodate unplanned events and necessities.

Discussion

Biomedical informatics and data science internships have typically been siloed, requiring students to choose one sector over another, often pitting academia against industry. Yet, biomedical informatics is an interdisciplinary, fluid field where knowledge transformation and innovation often occur through the intersection of academia and industry. There is a need for future informaticians and data scientists to possess this important aspect of cross-sector. Through the IBM-VUMC Academic Partnership, we created a novel internship that introduced undergraduate students to the biomedical informatics and data science fields through academic and industry experiences. Through this internship, interns gained interdisciplinary training and research skills applicable in both academia and industry; professional skills including scientific writing, public speaking, team science, and leadership; and networking opportunities to help them envision their place in the future of biomedical informatics. The development of an academic-industry internship in the biomedical informatics and data science fields was a worthy and successful investment in training tomorrow’s scientists and informaticians. Future efforts include a formal program evaluation.

Acknowledgment. The authors wish to thank Shilo Anders, PhD, Hu T. Huang, PhD, Allison B. McCoy, PhD FAMIA, Molly McKillop, PhD, Travis Osterman, DO MS, Rubina F. Rizvi, MD PhD, S. Trent Rosenbloom, MD MPH FACMI FAMIA, and Dilhan Weeraratne, PhD, for serving as mentors.

Learning Objectives

At completion of this activity, participants will be able to (1) State the value of joint academic-industry internships and (2) Understand the processes in developing a successful academic-industry informatics and data internship.

References

How Will AMIA Lead? An Environmental Scan of Diversity, Equity, and Inclusion Activities within the Healthcare and Health IT Space

Tiffani J. Bright, PhD, FACMI1, Carolyn Petersen, MS, MBI, FAMIA2, Karen Wang, MD, MHS3, Clair Kronk, PhD4, Patricia C. Dykes, PhD, RN, FAAN, FACMI5
1IBM Watson Health, Cambridge, MA; 2Mayo Clinic, Rochester, MN; 3Yale School of Medicine, New Haven, CT; 4University of Cincinnati School of Medicine, Cincinnati, OH; 5Brigham and Women’s Hospital, Harvard Medical School, Boston, MA

Introduction
In June 2020, the AMIA Board of Directors established the Diversity, Equity, and Inclusion (DEI) Task Force (now Committee) to advise AMIA on specific, actionable steps addressing matters of racial diversity, equity, and inclusion for strategic planning. An immediate charge of the DEI Task Force was to conduct an environmental scan of DEI activities in similar associations and societies. Environmental scans are efficient for assessing how organizations have acquired data and information to establish actions but also set strategic direction1. Objectives were to: 1) identify activities that other associations in the healthcare community were undertaking to advance DEI within their membership and the broader community; and 2) build a knowledge base on which to draw for strategic planning.

Methods
Web searches were undertaken to find organizations and societies serving health care and health information technology professionals in September 2020. Each website was searched for relevant results using the terms “diversity,” “equity,” “inclusion,” and “DEI.” Items that resulted in direct policy/position statements and/or activities were included; items such as press releases were excluded. During November and December 2020, content published on these organizations’ websites was reviewed and DEI-related policies/position statements and activities were captured. Policy/position statements and activities were grouped by type and purpose (e.g., policies applying to membership, professional development) and counted as a percentage of the number of organizations scanned.

Results
The web searches identified 30 organizations: 4 health IT organizations, 1 non-profit academic research organization, and 25 healthcare professional organizations. Policy/position statements from 5 of these centered on Black, Indigenous and People of Color physicians; 2 served physicians with diverse abilities; and 1 focused on health equity for LGBTQIA+ and sexual and gender marginalized health professionals. Across the organizations, 17% had no policy/position or activity, 47% had 1–5 policies/positions and/or activities, 23% had 6–8 policies/positions and/or Activities, and 13% had at least 10 policies/positions and/or activities. The organizations surveyed in this environmental scan published policies/position statements addressing 8 topical areas (Figure 1). Table 1 presents the most common policy/position statements.

<table>
<thead>
<tr>
<th>Policy/Position Type</th>
<th>%</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diversity</td>
<td>53%</td>
<td>Diversity in Health Services Research</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Establishing a Culturally Competent Masters and Doctorally Prepared Nursing Workforce</td>
</tr>
<tr>
<td>COVID-19 and Race</td>
<td>33%</td>
<td>Statement Urging HHS to Collect, Analyze, and Release COVID-19 Data on Mortality by Race &amp; Ethnicity</td>
</tr>
<tr>
<td></td>
<td></td>
<td>COVID-19 Treatments Must Work for Communities of Color</td>
</tr>
<tr>
<td>Racism</td>
<td>20%</td>
<td>Statement on Addressing Racism Through Collective Action in Ob/Gyn</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Addressing and Eliminating Racism at the AAMC and Beyond</td>
</tr>
<tr>
<td>Discrimination</td>
<td>20%</td>
<td>Discrimination in Membership Evaluation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Statement on Harassment, Bullying, and Discrimination</td>
</tr>
</tbody>
</table>

Figure 1. DEI Policy/Position Statements
Eleven types of activities were used to promote DEI across the organizations (Figure 2). Table 2 presents examples of the most prevalent types of DEI activities.

**Table 2. Most Common DEI Activities**

<table>
<thead>
<tr>
<th>Activity Type</th>
<th>%</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Professional Education</td>
<td>43%</td>
<td>Certificate in Diversity Management&lt;br&gt;Mid-career minority faculty leadership seminar</td>
</tr>
<tr>
<td>Committee/Task Force</td>
<td>40%</td>
<td>DEI Committee/commission&lt;br&gt;Advisory committee on LGBTQ issues</td>
</tr>
<tr>
<td>Research/Survey</td>
<td>37%</td>
<td>Diversity in Medicine: Facts and Figures 2019&lt;br&gt;Accessibility, Inclusion, and Action in Medical Education: Lived Experiences of Learners and Physicians with Disabilities</td>
</tr>
<tr>
<td>Professional Groups</td>
<td>33%</td>
<td>National Hispanic Leadership Agenda&lt;br&gt;Global Health Equity Network African American Community</td>
</tr>
</tbody>
</table>

**Figure 2. DEI Activities**

**Limitations**

The authors focused on organizations that were most similar to AMIA to gain a perspective specific to healthcare; this was informed by the DEI Task Force Charge. Future work may include examining examples of DEI efforts outside the healthcare domain that can be adopted for our purposes. Additionally, organizational materials reviewed were only available through the website, which may have resulted in missing some policy/position statements and/or activities.

**Conclusion**

This environmental scan found common DEI activities AMIA could implement to create a culture of inclusion and belonging for members and within the informatics profession. The findings demonstrated that there is no singular activity pursued by organizations, rather, multipronged strategies were the norm across these diverse organization. We also observed commonalities in activities across organizations that need to be explored further by engaging stakeholders from these organizations to learn about their membership outreach strategies for building an inclusive community. The widespread activities across many similar organizations reaffirms the urgency and importance of sustained resource-committed AMIA DEI efforts. As a next step, the DEI Committee is combining these findings with results from the 2021 AMIA DEI listening sessions to inform and prioritize the AMIA Board-approved recommendations. The DEI Committee and subcommittees will also be discussing organizational outreach for additional information and partnership opportunities, and additional policies/positions, and activities to transform AMIA into the professional organization that reflects society and the communities we serve, enabling members and the broader informatics workforce to influence and lead the transformation of healthcare.

**Acknowledgments.** The authors thank the DEI Task Force members and AMIA staff for sharing their insights.

**Learning Objectives**

At completion of this activity, participants will be able to (1) Understand how the AMIA Diversity, Equity, and Inclusion Task Force addressed the DEI Charge given by the AMIA Board to conduct an environmental scan.

**References**

To Deep or Not to Deep: Comparison of Traditional and Deep Learning Models in Disease Prediction from Electronic Health Records

Alon Brutzkus, MSc., Guy Amit, PhD.
KI Research Institute, Kfar Malal, Israel

Introduction

Prediction models based on deep neural networks (DNN) have become the state of the art technology for computer vision, natural language processing and speech recognition. Although their applicability for structured data from electronic health records (EHR) have been demonstrated in multiple studies [1], the reported performance gain over traditional linear or tree-based models varies considerably and may depend on specific characteristics of the data or on the study design. Self Attention with Reverse Distillation (SARD), a DNN architecture based on transformers has been recently reported to outperform logistic regression in clinical prediction tasks using claims data [2]. We aimed to evaluate SARD on a disease prediction task and compare its performance to gradient boosted decision trees, which are widely-used traditional models for structured data. For our predicted outcome we selected postpartum depression (PPD), one of the most common complications of childbearing, estimated to affect 10-15% of mothers worldwide [3]. Early identification of women at risk of PPD is clinically important, as timely interventions may improve the health outcomes of both mothers and infants [4]. PPD is associated with a variety of physiological, psychological and socio-demographic risk factors and is therefore an appropriate use case for EHR-based disease prediction.

Methods

Data: We analyzed primary care EHR data from from IQVIA Medical Research Data (IMRD), incorporating data from The Health Improvement Network (THIN, a Cegedim database). This dataset contains primary care records of over 12.5 million patients, covers approximately 5% of the UK population, and is representative of the population in terms of demographics and major condition prevalence [5]. The data includes patient demographics, medical diagnoses, drug prescriptions and lab tests, which were transformed into the OMOP common data model [6].

Population: Our cohort included women who gave live birth between 07/2000 and 03/2018, at age 15 to 50 years, and have been continuously observed at the same clinic for at least one year prior to their pregnancy and one year following giving birth. PPD outcome was defined by having either depression diagnosis or new treatment for depression during the year after childbirth. The group of women without PPD was downsized to obtain a balanced dataset.

Feature extraction: For the traditional models, we extracted, for each patient i, a set Ci of all medication, condition and procedure codes that were recorded in their medical file before the time of delivery. In total we obtained a set C of 25,317 codes. We created a feature vector \( f_i = [f_i(W^{(all)}), f_i(W^{(730)}), f_i(W^{(365)}), f_i(W^{(180)}), f_i(W^{(30)})] \in \{0, 1\}^{5|C|} \) where \( W^{(l)} \) is a time window of l days before the delivery, and \( f_i(W^{(l)}) \) is a one hot vector of dimension \( |C| \) where the element corresponding to \( c \in C \) is 1 if and only if \( c \) appears in a visit in the time window \( W^{(l)} \). Window \( W^{(all)} \) refers to the entire observation period before the delivery. The SARD algorithm learns nonlinear visit representation via embedding of the concepts in each visit, followed by applying a transformer to the sequence of visits (see [2] for details).

Models: We experimented with the following machine learning models: (1) L1-regularized windowed logistic regression (LR) with regularization parameter chosen from [2, 0.2, 0.02] based on the validation set; (2) Extreme gradient boosting (XGBoost [7]) with 140 estimators, maximum depth of 5 and learning rate = 0.1; (3) SARD [2] with minor adjustments to the available implementation. The \( \alpha \) coefficient of the reverse distillation loss was chosen from [0.0, 0.2] based on the validation set.

Experiments: We trained all models and evaluated their performance using 5-fold cross validation, where in each iteration 20% of the data was used as a test set, and a validation of the same size was randomly chosen from the training set. We calculated the mean and 95% confidence intervals (CI) of the area under the receiver operator characteristics [https://github.com/clinicalml/omop-learn](https://github.com/clinicalml/omop-learn)
curve (AUROC) for each model, as well as for an ensemble of the XGBoost and SARD models, using bootstrapping on the pooled prediction scores.

Results
The cohort included 49,911 women with mean age of 30.1±6.1 years, 20,921 of them with PPD outcome and the remaining 28,990 without PPD. SARD and XGBoost achieved similar AUROC of 0.779 [CI 0.775 - 0.783] and 0.782 [0.778 - 0.786], respectively, slightly improving LR (AUROC 0.769 [0.765 - 0.774]). An equally-weighted ensemble of SARD and XGBoost provided an additional small improvement (AUROC 0.786 [0.782 - 0.79]) (Fig 1).

Discussion
Although deep learning algorithms for healthcare data, such as SARD have been reported to achieve state-of-the-art results on some clinical prediction tasks, our experiments indicate that their advantage in the general case is questionable. On our PPD prediction task, gradient boosted trees achieved slightly better performance than SARD, with approximately x100 faster execution time. The improvement achieved by the ensemble model implies that tree-based and deep classifiers may be combined to provide optimal performance. In-depth exploration of the differences between the models and their appropriateness for various types of problems and data is a future extension of this study.

References
Linguistic Indicators of Behavioral Activation in Text-Based Therapy Sessions Anticipate Changes in Depression Symptomatology

Hannah A. Burkhardt, BS¹, George S. Alexopoulos, MD², Michael D. Pullmann, PhD¹, Thomas D. Hull, PhD³, Patricia A. Areán, PhD¹, Trevor Cohen, MBChB, PhD, FACMI¹
¹University of Washington, Seattle, WA; ²Weill Cornell Medicine, White Plains, NY; ³Talkspace, New York, NY

Introduction
Behavioral Activation (BA) is rooted in the behavioral theory of depression, which states that increased exposure to meaningful, rewarding activities is a critical factor in depression improvement¹. Previous work has shown that depressed and non-depressed individuals may use language differently and that automated tools can detect these differences. The increasing use of online, chat-based mental health counseling presents an unparalleled resource for longitudinal analysis of the language of patients with depression, with the potential to illuminate the role of reward exposure in recovery. Automatic assessment of constructs relevant to BA can facilitate research into these therapies, and guide their application in practice. We investigated how linguistic indicators of planning and participation in enjoyable activities identified in text-based counseling sessions relate to depression symptomatology over time.

Methods
Using messaging therapy logs from previous research² (>10,000 participants, >165 million total words), we developed a lexicon of words collectively capturing the construct of activation as used in BA. Lists of seed terms were constructed for each of the 7 items in the Activation subscale of the BADS questionnaire (examples shown in Table 1), expanded using distributional semantics methods, and iteratively refined with input from author G.A., a clinician with extensive experience in BA approaches. We expanded the seed sets of terms by adding the 30 most related terms as determined by cosine similarity between the terms’ vector representations in a space of 100-dimensional word embeddings trained using the skipgram-with-negative-sampling algorithm³ as implemented in the open source Semantic Vectors software package. This process yielded 7 subconstruct lexicons and one overall BA lexicon.

<table>
<thead>
<tr>
<th>Item</th>
<th>Derived seed terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>I am content with the amount and types of things I did. (satisfaction)</td>
<td>accomplish, achieve, satisfied, enjoy, content, accomplishment, love, proud, inspired, inspiring, enthuse, affirm</td>
</tr>
<tr>
<td>I engaged in a wide and diverse array of activities. (breadth)</td>
<td>activity, active, participate, involved, event, powerlifting, watercoloring, exercise, sport, basketball, restaurant, hobby</td>
</tr>
</tbody>
</table>

Table 1. Seed terms derived by the authors from the individual questions on the “Activation” subscale of the BADS. The name we assigned for each item (for brevity) is shown in parentheses.

Using Linguistic Inquiry and Word Count (LIWC)⁴, we then analyzed the language used in online therapy chat logs with respect to each of the 8 lexicons, as well as the established depression-related markers provided by LIWC, namely first-person singular pronoun (“I”, “me”, etc.) usage, first-person plural pronoun (“we”, “our”, etc.), and emotional tone (the balance between positive and negative affect). LIWC estimates scores for linguistic markers from the frequencies with which related words are used in text. The association of each of these linguistic markers with PHQ scores (depression severity), which were collected in regular intervals for all study participants, as well as the longitudinal development of depression scores over the treatment period (patient trajectory) was assessed. The patient trajectory groups (“improving” (n=3,211) and “non-improving” (n=3,546)) were identified on the basis of latent growth modeling results from previous research². Additionally, we used mixed effects linear regression modeling to determine the relative utility of each marker for predicting PHQ scores and patient trajectories.

Results
The level of activation (average percentage of words belonging to the overall activation construct) across the 10,711 baseline chat logs was 3.66 (SD 0.89) and varied significantly with the depression symptom severity category (Figure 1), as did the LIWC emotional tone measure and the LIWC pronoun measures. Changes in BA with depression severity are consistent with expectations, with activation rising as depression symptoms decrease.

The variance in the overall PHQ score explained (R²) by models fitted to different sets of variables is shown in Figure 2. BA constructs collectively explain more of the variance than the three LIWC variables (“we”, “I”, “tone”) (0.694 vs. 0.629). The combination of the novel and baseline variables explains more variance than either independently (R²...
Tone added more to satisfaction, breadth, accomplishment, and effort than to decisions, long-term planning, and daily structure, suggesting that these variable clusters may capture independent clinical dimensions.

Figure 1. Mean (95% CI) of each LIWC measure by depression symptom severity category at baseline: minimal (PHQ≤4, n=393), mild (PHQ=5-9, n=1,865), moderate (PHQ=10-14, n=4,109), moderately severe (PHQ=15-19, n=3,002), severe (PHQ≥20, n=1,331).

Mixed-effects linear regression confirmed that depression symptoms as measured by PHQ-9 scores improve significantly more for patients in the “improving” group than for patients in the “non-improving” group: a 1-week change was, on average, associated with a 0.37 (95% CI: 0.34-0.40) point improvement in PHQ in the “non-improvement” group vs. 0.71 (95% CI: 0.68, 0.74) in the “improvement” group. Similarly, each of the linguistic indicators showed significantly different amounts of change over time between groups.

Discussion
This work makes several key contributions. First, we devised a computational method to automatically assess theoretical constructs of BA from patient language, mediating measurement of the extent to which chat-based BA therapy is working as intended in large-scale clinical trials. Second, building on prior work demonstrating that activation has a close relationship with depression scores, we demonstrated that this new metric varies with depression symptom severity. Third, we validated established linguistic markers of depression at an unprecedented scale, using a large corpus of naturally occurring language collected as part of psychotherapy sessions. Fourth, we showed that both the well-established LIWC measures as well as the novel BA measures have utility in predicting longitudinal patient trajectories, with implications for the customization of clinical care. Finally, we demonstrated that our metrics of the individual subconstructs of BA capture distinct dimensions of the underlying mechanisms and may lend themselves to unique clinical insights. This work therefore enables further work in automated diagnosis and assessment of depression, as well as refinement of BA psychotherapeutic strategies.

Acknowledgements
This work supported by NLM Training Grant 5T15 LM007442-19 (HAB), National Institutes of Mental Health (NIMH) grants R01 MH102252 & P50 MH113838 (GSA), and NIMH grant P50 MH115837 [MDP, PAA].

References
Obesity Classifier Performance Disparities by Demographic Subgroup

Elizabeth A. Campbell, MS, MSPH1,2, Saurav Bose, MS2, Aaron J. Masino, ME, PhD2,3
1Drexel University, Philadelphia, PA; 2Children’s Hospital of Philadelphia, Philadelphia, PA; 3University of Pennsylvania Perelman School of Medicine, Philadelphia, PA

Abstract

Electronic Health Records are important data sources that may be used to develop machine learning models in healthcare research. In this work, classification models are developed to recognize pediatric obesity using temporal condition patterns obtained from patient EHR data. Classifier performance is evaluated by demographic subgroups to identify potential performance disparities. XG Boost and Logistic Regression were the best performing algorithms; algorithms performed similarly across subgroups but tended to be more accurate for minority populations.

Introduction

Childhood obesity is a major public health challenge in the United States that disproportionately impacts racial/ethnic minority groups and socioeconomically disadvantaged households. Electronic Health Records (EHRs) assist childhood obesity diagnosis, treatment, and surveillance at the clinical and population level. While machine learning models trained on EHR data are increasingly used in predictive medicine, there has been limited research on utilizing these methods to predict childhood obesity or potential disparities in classifier performance among vulnerable patient subpopulations. The following study develops a classification model to identify obesity among pediatric patients using clinical diagnostic patterns in their medical history and analyzes model performance by demographic subgroups.

Methods

EHR data was obtained from the Pediatric Big Data (PBD) resource at the Children’s Hospital of Philadelphia (CHOP). Patients in this study were from a retrospective cohort of newly obese patients and matched control patients with a healthy BMI (n = 99,388) identified in a previous study. The sequential pattern mining algorithm SPADE was applied to this cohort in the previous study to identify common temporal condition patterns that surround pediatric obesity incidence. From that study, we retained case-control pairs where patients had BMI measurements at all visits, leading to a population of n = 9,686 for this study. Seventy temporal condition patterns were considered for this study. Each pattern was considered separately as a feature; patients were assigned a binary value of 0 or 1 (indicating that a patient did or did not have a record of a given temporal condition pattern). The Institutional Review Board at the Children’s Hospital of Philadelphia (CHOP) approved this research study and waived the requirement for consent.

We trained four machine learning algorithms (Logistic Regression, Random Forest, XG Boost, and Neural Net) to classify cases and controls as obesity positive or negative, and optimized hyperparameter settings through a bootstrapping methodology. We randomly shuffled the data and split it into training and validation folds in a stratified fashion relative to class balance. We trained each model with all hyperparameter settings on the training fold and evaluated its Area under the Receiver Operating Curve (AUC-ROC) on the validation fold. We repeated the process 200 times to obtain 200 validation AUC-ROCs for each hyperparameter setting for each model, then selected the hyperparameter combination with the highest mean AUC-ROC for a given model class. We implemented all algorithms using the Scikit-learn library in Python 3. We calculated the mean and standard deviation (SD) AUC-ROC values for the total study population and demographic subgroups for each algorithm.

Results

The study population is majority male (55.4%) and majority White (60.1%). African Americans (AA) are the second largest racial/ethnic group (25.3%). Approximately 1/3 of patients (32.1%) were enrolled in Medicaid at the time of their index visit. The case population is majority male (55.4%) and majority White (53.9%), but is comprised of a higher proportion of African Americans (31.7% vs. 25.3%) and Hispanic patients (4.4% vs. 3.5%) patients compared to the study population as a whole. Additionally, a greater proportion of case patients (38.1%) were enrolled in Medicaid (a proxy for low socioeconomic status) compared to the overall study population (32.1%).
Table 1. Mean(SD) AUC-ROC for Study Population and Demographic Subgroups by Classification Algorithm

<table>
<thead>
<tr>
<th></th>
<th>Logistic Regression</th>
<th>Random Forest</th>
<th>XG Boost</th>
<th>Neural Net</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total Study Population</strong></td>
<td>0.78 (0.01)</td>
<td>0.77 (0.01)</td>
<td>0.78 (0.01)</td>
<td>0.76 (0.01)</td>
</tr>
<tr>
<td><strong>Gender</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>0.78 (0.01)</td>
<td>0.77 (0.01)</td>
<td>0.78 (0.01)</td>
<td>0.76 (0.01)</td>
</tr>
<tr>
<td>Female</td>
<td>0.78 (0.01)</td>
<td>0.77 (0.01)</td>
<td>0.78 (0.01)</td>
<td>0.77 (0.01)</td>
</tr>
<tr>
<td><strong>Race</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asian</td>
<td>0.76 (0.05)</td>
<td>0.76 (0.05)</td>
<td>0.77 (0.05)</td>
<td>0.74 (0.06)</td>
</tr>
<tr>
<td>Black/African American</td>
<td>0.79 (0.01)</td>
<td>0.79 (0.01)</td>
<td>0.79 (0.02)</td>
<td>0.78 (0.02)</td>
</tr>
<tr>
<td>White</td>
<td>0.75 (0.04)</td>
<td>0.75 (0.04)</td>
<td>0.76 (0.01)</td>
<td>0.74 (0.01)</td>
</tr>
<tr>
<td>Hispanic</td>
<td>0.75 (0.04)</td>
<td>0.75 (0.04)</td>
<td>0.77 (0.04)</td>
<td>0.74 (0.04)</td>
</tr>
<tr>
<td>Multiple Race</td>
<td>0.73 (0.08)</td>
<td>0.73 (0.08)</td>
<td>0.76 (0.08)</td>
<td>0.75 (0.07)</td>
</tr>
<tr>
<td>Unknown</td>
<td>0.73 (0.03)</td>
<td>0.73 (0.03)</td>
<td>0.73 (0.04)</td>
<td>0.72 (0.04)</td>
</tr>
<tr>
<td><strong>Medicaid Enrollment</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medicaid/CHIP</td>
<td>0.80 (0.01)</td>
<td>0.79 (0.01)</td>
<td>0.80 (0.01)</td>
<td>0.79 (0.01)</td>
</tr>
<tr>
<td>Not Enrolled in Medicaid/CHIP</td>
<td>0.76 (0.01)</td>
<td>0.75 (0.01)</td>
<td>0.76 (0.01)</td>
<td>0.74 (0.01)</td>
</tr>
<tr>
<td><strong>Age at index visit</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2-4 years</td>
<td>0.76 (0.02)</td>
<td>0.75 (0.01)</td>
<td>0.76 (0.02)</td>
<td>0.75 (0.01)</td>
</tr>
<tr>
<td>5-11 years</td>
<td>0.80 (0.01)</td>
<td>0.79 (0.01)</td>
<td>0.80 (0.01)</td>
<td>0.79 (0.01)</td>
</tr>
<tr>
<td>12-18 years</td>
<td>0.75 (0.02)</td>
<td>0.75 (0.02)</td>
<td>0.76 (0.01)</td>
<td>0.75 (0.02)</td>
</tr>
</tbody>
</table>

Mean AUC-ROC values were consistent across algorithms, ranging from 0.72-0.80. XG Boost and Logistic regression tended to perform the best on the full study population and when evaluated by demographic subgroups. Some evidence of bias was identified, although this was through the models performing better for minority subgroups. The highest mean AUC-ROC were observed among AA patients, patients enrolled in Medicaid, and patients ages 5-11 years.

**Discussion**

In this study, classifiers were trained to identify pediatric patients as obese or not obese using temporal condition patterns previously found to be associated with obesity incidence. Model performance was evaluated for the total population and by demographic classes. Surprisingly, the models tended to perform best on minority groups, including AA patients and patients enrolled in Medicaid, although tests of statistical significance were not performed. Our findings show that bias in complex machine learning models can manifest in varying ways. Prior work has shown that bias in machine learning models has occurred against minorities due to an under-representation in data. However, possible biases need to be examined carefully. We hypothesize that our models performed better on under-represented groups because the features more strongly associated with obesity were more common among minority patients. These findings can be incorporated into future research to develop a thorough analytical approach to identify and mitigate bias that may arise from features and within datasets when developing more equitable machine learning models.

**References**

Clinical Decision Support with a Comprehensive in-EHR Patient Tracking System to Improve Inpatient Genetic Testing Follow Up

Ian M. Campbell, MD PhD1,2; Morgan L. McManus, MS2; Fred C. Cusick, BS1; David C. Junod, BS1; Sarah E. Sheppard, MD PhD2; Eli M. Lourie, MD, MBA, FAMIA1,2; Eric D. Shelov MD, FAMIA1; Anthony A. Luberti, MD, FAMIA1

1Department of Biomedical and Health Informatics, Children’s Hospital of Philadelphia, Philadelphia, PA; 2Division of Clinical Genetics, Children’s Hospital of Philadelphia, Philadelphia, PA; 3Department of Pediatrics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA

Introduction

Genetic testing can unlock the promise of precision medicine, potentially informing prognosis, management, surveillance, and recurrence risk. However, to be useful, the testing must be reviewed and interpreted by professionals who can implement care based on the knowledge gained and inform the patient or their family members of its meaning. The inpatient pediatric clinical genetics evaluation workflow involves the primary attending physician requesting consultation of a geneticist (Figure 1). The geneticist evaluates the patient and makes recommendations for genetic testing. It is the responsibility of the primary team to send samples to the hospital or outside reference laboratory for analysis. The laboratory issues a test report which is in turn entered into the patient’s electronic health record (EHR). Because the genetic testing takes weeks or even months to complete, the patient is frequently discharged before the results become available. Ideally, the clinical genetics team would become aware of the results and interpret them based on patient and family history and document their findings in the EHR. In an attempt to facilitate this, the health care team has historically developed ad hoc, sometimes intensely personal systems of Microsoft Office® documents and paper records, a practice not uncommon in other medical contexts1. Because of the complex nature of the workflow, the timescales involved, and sociotechnical issues within the EHR, results may go unnoticed. In an attempt to improve genetic testing result interpretation and follow up documentation, we developed a clinical decision support mechanism using a comprehensive in-EHR tracking system.

Methods

We analyzed the volume, genetic testing performed, result access logs, and follow up documentation of inpatient clinical genetics consultation to determine the scope of the problem and facilitate outcomes evaluation. Data visualization and statistical analyses, including multivariate logistic regression, were performed in R. Informed by the data analysis, we developed a tracking system with two main components. Part 1 is a data capture instrument which is available to clinicians in note templates and from patient lists (Figure 2). The capture form is optimized to reduce double documentation and provides customized suggestions but gives the clinician an opportunity to alter the parameters based on clinical judgement. Part 2 is a management dashboard which gives clinicians an overview of patients with outstanding tasks, provides easy access to information required to make decisions, and facilitates communication with the primary team and the patient’s family (Figure 3).

Results

We analyzed the outcomes of 772 inpatient genetics consultations during the 1-year period prior to implementation of the tracking system. 58% of patients had no subsequent documentation from a provider or staff member of clinical genetics. Of those patients, 78% (or 45% of the total), had genetic testing performed after their consultation but had no follow up documented by the genetics team. It is potentially possible that some of these patients

Figure 1 Clinical genetics inpatient consultation

Figure 2 Data capture instrument for the in-EHR tracking system
had results reviewed and documented by their primary team. We developed a multivariate logistic regression model to predict documentation of follow up based on completion of genetic testing, patient race and ethnicity, preferred language, payor type, and socioeconomic status of their home zip code. Unsurprisingly, we found that performing genetic testing and especially exome sequencing, the most complicated and expensive genetic testing, was strongly positively associated with follow up documentation ($p<0.002$, multivariate logistic regression). Notably, the patient having government insurance was significantly negatively associated with follow up documentation ($p=0.039$, multivariate logistic regression), suggesting a potential socioeconomic influence.

Uptake of the tracking system by the user base has been swift. The system was fully implemented on December 1st, 2020. In the first 6 months, 333 of 373 potentially eligible inpatient consults (89.3%) have been entered into the system. The system facilitated multiple diagnoses that would have otherwise been missed, including a case of 22q11.2 deletion syndrome because genetic testing was inadvertently not sent by the primary management team.

Considering data collected in the 6 months between December 1st, 2020 and June 1st, 2021, 171 of 275 (62.2%) inpatient consults who had genetic testing and were tracked by the system had follow up documentation within 60 days. This is a significantly greater proportion than those not tracked by the system over the previous year, 276 of 530 (52.1%, $p<0.005$, Two-sided Fisher’s Exact Test). Although documentation rates differed among individual attending physicians, use of the tracking system was associated with increased follow up documentation by all providers.

Using EHR access log data, we found that review of genetic results was completed by a clinical genetics provider within one month for a great majority of patients regardless of tracking system utilization (Figure 4). However, the system helped facilitate review of small subset of patients that may otherwise have gone unnoticed ($p<0.001$, Kolmogorov-Smirnov test). Indeed, all results have thus far been reviewed within 3 months (max 73 days) with the tracking system, whereas 8.2% of results remain unaccessed by that time without the system.

**Discussion**

Overall, we have implemented an in-EHR tracking system to serve as a clinical decision support tool that facilitates genetic testing and follow-up. The tool was quickly adopted by the clinical team. Our data suggests that the ability to easily see an overview of outstanding care tasks may expedite review of results and documentation of interpretation and recommendations. We believe the same framework can easily be extended to other medical specialties and patient care problems.

**References**

Data-Driven Patterns in Protective Effects of Ibuprofen and Ketorolac on Hospitalized Covid-19 Patients

Rich Caruana, PhD¹, Benjamin J. Lengerich, PhD², Yin Aphinyanaphongs, MD, PhD³
¹Microsoft Research, Redmond, WA; ²Massachusetts Institute of Technology, Cambridge, MA; ³NYU Langone Health, New York, New York

Introduction
The impact of nonsteroidal anti-inflammatory drugs (NSAIDs) on patients with Covid-19 is unclear. Initial guidance was to avoid prescription of NSAIDs to Covid-19 patients [1]. However, more recent meta-analyses have suggested that there is no compelling evidence of a harmful effect of ibuprofen on Covid-19 patients [2, 3]. It is difficult to identify the effect of NSAIDs due to confounding: NSAID usage in Covid-19 patients may be correlated to more severe cases of Covid-19 which manifest with inflammatory symptoms [4]; on the other hand, medical intuition suggests that because NSAIDs are contraindicated for some severe comorbidities, they may be correlated with lower-risk patients. Other analysis has found no association between ibuprofen and clinical outcomes, even without deconfounding [5]. Moreover, a few observational analyses have indicated that a protective effect from outpatient NSAIDs [6] or Ibuprofen¹ is observed after correcting for underlying patient risk factors. Finally, there is some evidence of protective effects of NSAIDs on patients with respiratory illnesses [7], although similar clinical trials have not been completed for Covid-19. In this setting, we seek to use machine learning tools to estimate patient risk at time of admission from patient lab values and then estimate the additive effect of Ibuprofen and Ketorolac.

Methods
Our outcome is in-hospital mortality. We consider only treatments within 24 hours of hospital admission; this captures a total of 97 patients who were treated with either Ibuprofen or Ketorolac. Our control condition is not receiving either NSAID in the first 24 hours. Our analysis includes deidentified records from 3108 patients hospitalized from March to August 2020 for Covid-19, with an average mortality rate of 18.1%. Our dataset for analysis was selected from a larger set of 11080 total hospitalized patients who have lab-confirmed cases of Covid-19. To filter out patients who were hospitalized for reasons other than Covid-19, we excluded patients who have indicators of (1) pregnancy: outpatient prenatal vitamins, in-patient oxytocics, folic acid preparations; or (2) scheduled surgery: urinary tract radiopaque diagnostics, laxatives, general anesthetics, antiemetic/antivertigo agents, or antiparasitics. We also require that the patients have recorded temperature, age, BMI, and Admission Day, and did not die within six hours of admission.

To correct for confounding, we use a two-stage machine learning procedure to estimate the adjusted risk ratio (ARR) [8], which can be interpreted as an odds ratio after correcting for patient mortality risk at admission. First, we take half of the sample and train a generalized additive model (GAM) [9, 10] to predict mortality risk at patient admission. We use tree-based GAMs [10] implemented in InterpretML². Next, in our held-out sample, we compute the excess risk after accounting for the model predictions of patient risk. To compute the effect of treatments, we compare the excess risk in treated patients with the excess risk in the control patients, and summarize these effects with the adjusted risk ratio (ARR) and adjusted risk difference (ARD) [8]. We use bootstrap to compute standard errors and p-values.

Results
The risk model achieves an ROC of 0.912 ± 0.001 and an F1-score of 0.598 ± 0.002 on held-out patients, significantly outperforming logistic regression which achieves an ROC of 0.859 ± 0.001 and F1-score of 0.455 ± 0.002. The most important features to this risk model are: Temperature, Age, Admission Day, Calcium, and Charlson score (Fig. 1). Estimated effects of Ibuprofen and Ketorolac are shown in Table 1. The observed mortality rate of patients treated with Ibuprofen or Ketorolac is lower than would be expected based on patient risk factors (comorbidities, demographics, and lab tests) at admission. For Ketorolac, the evidence is sufficient to reject a null hypothesis of no effect at p=0.05.

---
²https://github.com/interpretml/interpret
<table>
<thead>
<tr>
<th>Drug</th>
<th>N</th>
<th>Obs. Mort. (%)</th>
<th>Exp. Mort (%)</th>
<th>ARD (95% CI)</th>
<th>ARR (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ibuprofen</td>
<td>55</td>
<td>5.17 ± 3.06</td>
<td>7.93 ± 4.77</td>
<td>−2.82(−6.41, 1.80)</td>
<td>0.65(0.45, 1.54)</td>
</tr>
<tr>
<td>Ketorolac</td>
<td>50</td>
<td>2.46 ± 1.98</td>
<td>6.78 ± 3.81</td>
<td>−4.29(−7.95, −1.08)</td>
<td>0.36(0.24, 0.69)</td>
</tr>
<tr>
<td>Ibuprofen or Ketorolac</td>
<td>97</td>
<td>4.42 ± 2.02</td>
<td>7.97 ± 3.21</td>
<td>−3.50(−6.51, −0.74)</td>
<td>0.56(0.40, 0.86)</td>
</tr>
</tbody>
</table>

**Table 1:** Observed and expected rates of mortality for patients treated with Ibuprofen or Ketorolac.

**Figure 1:** Additive effects of Temperature and Age. Shaded regions are 90% CIs. Yellow ticks indicate 10 patients.

**Discussion**

These results add to the growing body of evidence that NSAIDs, and in particular Ibuprofen or Ketorolac, may not have detrimental effects on patients hospitalized with Covid-19. While these medications were correlated with a younger subset of the hospitalized patients and later treatment dates, the protective effects remain after correcting for confounding. This suggests that further study, including randomized control trials, should be performed to better understand any potential benefits of Ibuprofen and Ketorolac on patients hospitalized with Covid-19.

**References**


Deep learning detection of reticular pseudodrusen using multi-modal, multi-task, and multi-attention mechanisms: towards automated and accessible classification of age-related macular degeneration

Qingyu Chen, PhD1*, Tiarnan D. L. Keenan, BM BCh, PhD2*, Emily Y. Chew, MD2†, Zhiyong Lu, PhD1†, for the AREDS2 Deep Learning Research Group3

1. National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health (NIH), Bethesda, MD, USA
2. Division of Epidemiology and Clinical Applications, National Eye Institute, National Institutes of Health, Bethesda, MD, USA
3. See appendix

* These authors contributed equally to this work
† To whom correspondence should be addressed: echew@nei.nih.gov; zhiyong.lu@nih.gov

Abstract

Age-related macular degeneration (AMD) is the leading cause of visual loss in developed countries. Reticular pseudodrusen (RPD) are a critical disease feature of AMD (associated with increased risk of progression to advanced AMD). We propose a deep learning framework to detect RPD automatically. It achieved an Area Under the Curve (AUROC) of ~0.9 and ~0.8 on fundus autofluorescence images and color fundus photographs, respectively. It also achieved an AUROC of ~0.9 on an external dataset.

Introduction

Age-related Macular Degeneration (AMD) is the leading cause of blindness in developed countries. By 2040, 288 million patients are projected to have AMD worldwide (1). Following these developments in retinal imaging, an important macular feature -- reticular pseudodrusen (RPD) -- is now recognized as a key AMD lesion (2, 3). RPD presence is strongly and independently associated with increased risk of progression to late AMD. While it is pressing to identify the presence of RPD, it is very poorly visible to human eyes on clinical examination for color fundus photography (CFP), the most widely used and accessible imaging modality in ophthalmology, even to trained experts at the reading center level (4). In contrast, only more recent retinal imaging techniques such as fundus autofluorescence (FAF) is effective to identify RPD, but it is not widely available in the clinical practice.

In this work, we propose a novel deep learning framework for the detection of RPD detect RPD presence accurately using CFP alone, FAF alone, or both, employing >8000 CFP-FAF image pairs obtained prospectively as part of the Age-Related Eye Disease Study 2 (5). The framework includes multi-modal (detection from single or multiple image modalities), multi-task (training different tasks simultaneously to improve generalizability), and multi attention (improving ensembled feature representation) operation (6). It has been the first study on the detection of RPD using multiple imaging modalities. Importantly, with the multi-task feature, the final models support all three image scenarios (CFP alone, FAF alone, or CFP-FAF pairs) as inputs; therefore, clinicians could use it according to the availability of imaging equipment.

Performance of the models was compared with 13 human retinal specialists; an external validation on patients from different populations (Rotterdam Study, Netherlands) was also performed. The models achieved area under receiver operating characteristic (AUROC) 0.832, 0.931, and 0.933 for CFP alone, FAF alone, and both, respectively. Its performance on CFP was very substantially superior to human retinal specialists (median F1-score 0.644 versus 0.350). The External validation also demonstrated high accuracy of the models; e.g., it had an AUROC of 0.965 for CFP alone. Overall, this work demonstrates the successful development, robust evaluation, and external validation of a novel deep learning framework that enables accessible, accurate, and automated AMD diagnosis and prognosis.

Data and methods

The primary dataset consisted of all AREDS2 images where a CFP-FAF pair was available, i.e., where a CFP and a corresponding FAF image (taken from the same eye at the same study visit) were available. The dataset is described with these CFP-FAF pairs as the imaging unit. The total number of images was 8487 (i.e. 8487 CFP, 8487 FAF images, and 8487 CFP-FAF image pairs). The dataset was split randomly into three sets: 70% for training, 10% for validation, and 20% for testing of the models.
The framework consists of three deep learning models: the CFP model, the FAF model, and the CFP-FAF model. The CFP model takes CFP images as its input and predicts RPD presence/absence as its output; the same idea applies to the FAF and CFP-FAF models. For the CFP model and FAF model, each has a CNN to extract features from the input image, followed by an attention module, followed by fully-connected layers, and an output layer, making the prediction. The CFP-FAF model has the same structure except that, instead of having its own CNN backbone, it receives the image features from both the CFP and the FAF models. The method is summarized in detail in (6).

The following metrics were used to evaluate the performance: F1-score, area under receiver operating characteristic (AUROC), sensitivity, specificity, Cohen’s kappa, accuracy, and precision. The model performance was further compared with the performance of 13 ophthalmologists (from ‘fellow level to ‘attending’ level).

Figure 1. Evaluation results of the models and the ophthalmologists for both CFP and FAF imaging scenarios (6).

Results and discussions

For RPD detection, the model achieved area under receiver operating characteristic (AUROC) 0.832, 0.931, and 0.933 for CFP alone, FAF alone, and both, respectively. As shown in Figure 1, its performance on CFP was very substantially superior to human retinal specialists (median F1-score 0.644 versus 0.350). We further performed an external validation (on Rotterdam Study, Netherlands). It also demonstrated high accuracy on CFP alone (AUROC 0.965). Finally, we evaluated the framework on detecting other AMD disease features: geographic atrophy and pigmentary abnormalities. It achieved an AUROC of 0.909 and 0.912, respectively, demonstrating its generalizability.

Overall, we believe that this framework demonstrates the potential for automated but accurate ascertainment of the full spectrum of AMD features from CFP alone. This is extremely valuable for improved AMD classification and risk prediction. Importantly, operation from CFP alone makes it accessible far beyond the small number of specialist centers in the developed world with access to multi-modal imaging and expert graders. Planned future work consists of additional external validation and prospective assessment in a clinical trial setting. This would take a step towards clinically applicable artificial intelligence systems.

References

Reproducibility in natural language processing: What we need to know

KB Cohen, PhD1, A Ripple, MLS2, A Ben Abacha, PhD2, O Bodenreider, MD, PhD2, O Hargraves, BA3, K Verspoor, PhD4, P Zweigenbaum, PhD5, D Demner-Fushman, MD, PhD2

1 U. Colorado School of Medicine, Denver, CO, USA; 2 Lister Hill National Center for Biomedical Communications, National Library of Medicine, NIH, USA; 3 U. Colorado, Boulder, CO, USA; 4 RMIT University, Australia; 5 LISN, CNRS, U. Paris-Saclay, France

Introduction

The philosopher of science Hans Radder put it well: in order to know whether or not an experiment has been reproduced, we first need to know what was actually done1. But, case studies have demonstrated that it is not always intuitively obvious what exactly needs to be reported about a natural language processing experiment2. Furthermore, work by Olorisade et al.3 has demonstrated that this is not always clear even for a relatively constrained subfield, such as text mining research. When we begin to think about the fundamental issue of generalizability in reproducibility, the question becomes even more complicated, and the answers probably much more nuanced4.

To address the basic question of what was actually done in this natural language processing experiment?, we take an approach based in the FAIR Principles for managing the products of scientific research5. The FAIR Principles propose that products of research, including publications and data, be Findable, Accessible, Interoperable, and Reusable. For publications, this suggests that a paper should be retrievable and analyzable by the semantics of its contents, and not only by search on the words with which that content is written. Although there are alternatives, it is a good choice because of its wide acceptance in biomedical research (see the reviews in6 and7; at the time of writing, the original paper has been cited over 4500 times).

The goals of retrievability and analyzability suggest an ontology or terminology to which concepts mentioned in papers can be normalized. Consequently, we propose here a two-part schema for representing a natural language processing paper as a collection of metadata. The first part models a paper itself. The approach to this is frame-based and is inspired by previous work in the domains of biomedical research publishing8–10 and of standards for the reporting of experiments11–13. The second part is an ontology of slot-fillers for that frame. Because community consensus is essential to the adoption of any such representation of scientific work5, our overall approach includes a significant amount of solicitation of feedback from a diverse cross-section of the natural language processing community. Additionally, we tested the coverage of the ontology using frequency-based methods applied to the language processing and text mining literature from two relatively distinct communities—biomedical text mining, and the Association for Computational Linguistics family of conferences.

Materials and Methods: The representation of a paper consists of a frame with slots for the following four items:

1. Topic: what is the paper primarily about?
2. Method: what was done?
3. Data: what kind of material was used?
4. Evaluation: how was the work evaluated, question answered, or hypothesis tested?

The slot fillers form an ontology structured by the typical relations, i.e. is-a and has-part, and a few novel ones.

In order to minimize subjectivity, the first draft of the ontology was constructed based on the indexes and tables of contents of popular language processing textbooks. Definitions were taken from open-source materials, including the primary literature and Wikipedia, and reviewed by a lexicographer. The overall model of papers, as well as the ontology of slot-fillers for describing them, was evaluated in two ways: by quantitative comparison to frequency and terminological analyses of the literature, and by solicitation of feedback from researchers in the field.

Quantitative evaluation: We analyzed over 9,000 PubMed-indexed natural language processing and text mining papers. The Sketch Engine terminology extraction tools14 generated a silver standard for evaluation of coverage.

Expert feedback: We did initial annotations of the complete sets of PubMed-indexed publications of several authors. We then met with them individually, and they corrected the metadata that we assigned to their papers.
Table 1: High-level annotations for four typical biomedical natural language processing papers.

<table>
<thead>
<tr>
<th>Topic</th>
<th>Method</th>
<th>Data</th>
<th>Evaluation</th>
<th>Paper title</th>
</tr>
</thead>
<tbody>
<tr>
<td>Named entity recog.</td>
<td>HMM</td>
<td>Journal articles</td>
<td>Shared task</td>
<td>BioC Task1A: Finding NEs with a stochastic tagger</td>
</tr>
<tr>
<td>Text classification</td>
<td>SVM</td>
<td>Clinic notes</td>
<td>Gold standard</td>
<td>Predicting pediatric epilepsy surgery candidates</td>
</tr>
<tr>
<td>Summarization</td>
<td>Rule-based</td>
<td>Journal articles</td>
<td>Gold standard</td>
<td>Finding GeneRIFs via Gene Ontology annotations</td>
</tr>
<tr>
<td>Corpus</td>
<td>Distribution</td>
<td>Journal articles</td>
<td>Hypothesis testing</td>
<td>Text in traditional and Open Access scientific journals</td>
</tr>
</tbody>
</table>

Results: The ontology currently contains 390 unique concepts and several relation types. Table 1 shows the high-level metadata for four typical papers. Meetings with individual authors led to improvements in the granularity of the representation. The initial overlap after manual filtering of terminology extraction errors was 48%. Most missing concepts were very domain-specific, such as electronic health record and biomedical text.

Conclusion: Thanks to the combination of methodologies, this work has resulted in an ontology for the representation of natural language processing papers that is both empirically supported by quantitative data, and vetted by members of the natural language processing community.

References

Examining the Sociotechnical Process of Clinical Photography to Encourage Photo Diversity in Medical Education

Benjamin Collins, M.D., M.A.¹, Pamela Pierce, M.L.S., M.S.¹, Linda Felver, Ph.D., R.N.¹
¹Oregon Health & Science University, Portland, Oregon

Introduction:

Smartphone technology has allowed clinical photography to play an increasingly significant role in medical education for learning to recognize visible diseases. It should be noted that skin tone alters the appearance of many conditions, making it essential to view conditions on a variety of skin tones. However, representation of different skin tones is lacking for photographs in many medical education resources. This inequity leads to poorer patient care and the perpetuation of health disparities. To facilitate acquiring diverse photographs for medical education, it is necessary to understand the sociotechnical process of clinical photography, including the technology, workflow, policies, and ethics involved. The objective of this study was to gain an understanding of the process for clinical photography at an academic medical center through interviews with faculty. This will serve as a foundation for developing a process to ethically and efficiently obtain clinical photographs of patients with diverse skin tones for medical education.

Methods:

The IRB at Oregon Health & Science University (OHSU) in Portland, Oregon approved this study. Twelve 30-minute video interviews were conducted via the virtual meeting platform, WebEx, with faculty knowledgeable of the clinical photography process. The number of interviews was set by reaching a saturation of new information. Participants were contacted by email for initial consent and scheduling of the interview. Consent was reaffirmed at the time of the interview. Interviewers used two versions of an interview guide depending on whether the interviewee actively took photographs or had knowledge of the process but did not actively take photographs. There was some leeway to deviate from the question guide to pursue further information or clarification. After completion of the interview, video recordings were securely stored in an OHSU approved cloud service then transcribed to a Microsoft Word Document. The interviewee’s identity was redacted from the transcript followed by discarding of the video recording with interviews thereafter identified by a randomly generated three-digit number to ensure confidentiality. Transcripts were analyzed in QDA Miner Lite using deductive themes based on the question guide as a foundation with inductive subthemes revealed when reviewing the transcript text within each theme. These codes were organized in Microsoft Excel for ease of review.

Results:

Twelve interviews were conducted with ten clinicians, one medical photographer, and one clinical administrator from the following clinical departments: Dermatology, Plastic Surgery, Internal Medicine, Emergency Medicine, Pediatrics, Ophthalmology, Vascular Surgery, and Endocrinology. Qualitative assessment of the interviews yielded five main themes based on the interview guide: technology, workflow, benefits, ethics, and diversity. Within these themes, there were 19 inductive codes used to build a chronological and technical workflow of the clinical photography process.

The workflow for clinical photography starts from the decision to take a photograph which is largely dictated by clinical role. Some roles take photographs multiple times a day and others may do so on a less than weekly basis. Also, clinicians who are less comfortable with the process may take photographs less frequently. Taking clinical photographs expressly for educational purposes was not common. They were typically taken to track the appearance of a condition over time. Once a decision is made to take a photo, consent is obtained from the patient. For clinical purposes, verbal consent is considered sufficient, but for other reasons, including education, written consent is deemed necessary. The written consent process was considered a significant burden on time. However, the interviewees recognized the importance of communicating the reason for the photo with the patient and following HIPAA guidelines. They were also aware that institutional policy existed, although sparse on the details. After consent is obtained, the photograph is taken. Most of the interviewees used personal smartphones to take photographs via the Epic Haiku app that bypasses storage in the phone and places the photo directly in the Epic EHR. There were concerns about the lower quality of these images due to limitations of the smartphone camera and data compression from using Haiku. Some interviewees also acknowledged that not all faculty use the Haiku app, leading to photographs stored on their personal smartphone, putting confidentiality at risk. Only the medical photographer spent time setting up for photographs. For clinicians, efficiency was more important than quality. After taking the photograph, viewing it requires entering the patient’s chart. Searching for photographs in the EHR to use for education was considered
problematic because they are stored in the patient’s chart, and being able to recall patients with photos of specific diseases is difficult. Additionally, Haiku does not have editing capability to remove identifiable features from a photo which would be necessary to use it for education, but this can be completed later on a computer. When discussing the diversity of clinical photographs in medical education, many of the interviewees referenced the lack of diversity in the population served by the institution. Interviewees also discussed the need for more diverse photographs in medical education from the perspective of racial diversity without specifically discussing diversity of skin tone.

Discussion:

The interviews revealed a number of important areas to address when aiming to obtain more clinical photographs with diverse skin tones for educational use. While interviewees displayed a clear understanding of the need for more diverse images, there was a lack of proficiency in this understanding. The discussion centered on race, which is a social construct, but skin tone varies significantly within races. Having a photograph of a lupus discoid rash on a white person and a Black person does not achieve the same effect for learning to recognize the rash as having photographs of the condition on multiple skin tones. The history of exploitation of persons of color must also be acknowledged in this pursuit of health justice. Persons of color should not need to be concerned about being exploited for education, and should feel welcome to participate. While the provision of good clinical care remains the priority, an academic medical center should be upfront about its role in medication education. This understanding could be improved by obtaining written consent from each patient so that not every educational photograph needs new written consent at the time of the encounter and can instead be verbally reaffirmed. This would also reduce the burden of time on clinicians. There would be exceptions to this such as photographs of faces, genital regions, physical abuse, or other images that need increased privacy. It would be beneficial to involve people of different backgrounds in a focus group to determine the best way to achieve a balance of encouraging participation in medical education and avoiding mistrust.

An ideal educational database for clinical photographs would be paired with a secure smartphone app, similar to Epic Haiku, that transmits the photograph straight into the database without being stored in the phone. This app would also have the ability to edit out identifiable features at the time of the photograph being taken and the option to add metadata by which the photograph can be organized when stored in the database. Photographs stored in the database would maintain a level of resolution suitable for use in education. To make the most of this database, clinicians should be trained on clinical photography which would increase the number of photographs taken. Overall, the sociotechnical process of clinical photography requires supportive policies, an efficient workflow with a reliable database, and trust from the community to improve the diversity in photographs for medical education. Important limitations of this study include that it was performed at one institution with use of one EHR. There is also selection bias from participants interested in clinical photography, influencing the frequency of photographs and adherence to institutional policies.

Funding:

This study was partially supported by a National Network of Libraries of Medicine Pacific Northwest Region Health Sciences Library Partnership Award. Developed resources reported in this publication are supported by the National Library of Medicine (NLM), National Institutes of Health (NIH) under cooperative agreement number UG4LM012343 with University of Washington. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

A Retrospective Analysis of Machine Learning Driven Antibiotic Selection in the Emergency Department

Conor K. Corbin, MD, Arhana Chattopadhyah, MD, Lillian Sung, MD, PhD, Amy Chang, MD, Stan Deresinski, MD, Jonathan H. Chen, MD, PhD

1Department of Biomedical Informatics Research, Stanford University, Stanford, California; 2Division of Haematology/Oncology, The Hospital for Sick Children, Toronto, Ontario, Canada; 3Medicine and Infectious Diseases, Stanford Medicine, Stanford, California

Introduction
Over 700,000 people die globally due to antibiotic resistant infection, and this figure is expected to reach 10 million by 2050. Over 50% of antibiotic use in the hospital is either unwarranted or sub-optimal, expediting growing resistance. Empiric antibiotic administration is often required because of rapid clinical deterioration in the event of true infection. However, appropriate antibiotic selection is difficult because selection must be made before knowing the result of cultures and antibiograms of isolated organisms. Clinicians must balance broader coverage to ensure activity against life-threatening infection and narrower coverage that adheres to antibiotic stewardship goals. We hypothesized that antibiotic selection could benefit from machine learning based clinical decision support. In this study we conducted the following: 1) trained machine learning models using electronic health record (EHR) data that estimate the probability that twelve distinct antibiotic therapies would cover identified isolates from individual patients, 2) retrospectively benchmarked an optimized allocation of antibiotics that uses the outputs of the machine learning models against both a random allocation of antibiotics and the observed clinician allocation, and 3) estimated the extent to which patient coverage decreases as fewer broad-spectrum antibiotics are used.

Methods
We used the Stanford Research Repository (STARR) clinical data warehouse to extract de-identified EHRs for patients who visited the Stanford or Valley Care emergency departments (EDs). The unit of observation was a patient ED admission. Analysis was restricted to admissions between 2009 and 2019. We included patients 18 years or older who were admitted to the hospital from the ED. We further restricted the analysis to admissions where at least one order for a blood, urine, cerebrospinal fluid (CSF), or fluid microbial culture, and at least one order for intravenous or intramuscular antibiotics were placed in the first 24 hours after presenting to the ED. Patients who had a record of receiving antibiotics in the two weeks leading up until the ED admission were excluded. The analysis was also restricted to patients with at least one positive microbial culture.

Binary machine learning classifiers were trained to predict the probability that a patient would be covered by twelve distinct and commonly administered empiric antibiotic selections using EHR data collected up until the point antibiotics were ordered. Features included patient co-morbidities, procedures, labs, medications, vital signs, and prior microbiology lab results. Features were represented as counts, and our final feature matrix contained 43,220 columns. Labels representing sensitive organisms were derived from the results of microbial cultures. The label was positive if all microbial organisms isolated from a patient were listed as susceptible to at least one of the antibiotics in the antibiotic regimen. Train (2009-2017) validation (2018), and test (2019) splits were made based on the years the admissions took place to best mimic potential distributional shift that occurs because models are trained and deployed on data from different time periods. We conducted a model selection procedure including four model classes (l1 and l2 regularized logistic regressions, random forests, and gradient boosted trees) and associated hyperparameters.

We used the out of sample probability estimates of the final model for each of the twelve prediction tasks as inputs to a linear programming-based optimization procedure that maximized predicted patient coverage over the set of observations in our test set with several constraints. Let $N_{test}$ be the number of ED admissions in our held out test set and $M = 12$ be the total number of antibiotic selections. Let $S \in R^{N_{test} \times M}$ be a matrix of binary variables indicating antibiotic selections, where $s_{ij} = 1$ represents that the $j$th antibiotic selection is allocated to the patient in the $i$th ED admission. Let

$$\text{maximize} \quad \sum_{i=1}^{N_{test}} \sum_{j=1}^{M} \phi_{ij} s_{ij}$$

$$\text{subject to} \quad \sum_{j=1}^{M} s_{ij} = 1 \quad i = 1, ..., N_{test}$$

$$\sum_{i=1}^{N_{test}} s_{ij} = K_j \quad j = 1, ..., M$$
\( \Phi \in R^{N_{\text{test}} \times M} \) be a matrix of probability estimates from our classifiers, specifically \( \phi_{ij} \) represents the predicted probability that the patient in the \( ith \) ED admission would be adequately covered by antibiotic selection \( j \). Finally, let \( K_j \) be the total number of times the \( jth \) antibiotic must be used over the set of our \( N_{\text{test}} \) ED admissions - the budget for antibiotic \( j \). We first chose the \( K_j \) parameters to be the number of times clinicians actually allocated the \( jth \) antibiotic selection. We later perturbed these parameters to estimate the extent at which patient coverage decreases when fewer broad-spectrum antibiotics are allocated. Our problem formulation is shown in Equation 1 and was implemented using the PuLP python package and solved with the CBC solver. Using these data we calculated the miss rate, defined as the fraction of patients who were assigned an antibiotic regimen that did not cover their infection.

**Results**

Our training, validation, and test sets had \( N_{\text{train}} = 5,804 \), \( N_{\text{val}} = 1,218 \), \( N_{\text{test}} = 1,320 \) observations respectively. The model class of the final model, positive class prevalence, average precision, and area under the receiver operating characteristic (AUROC) curve for each of the twelve binary classification tasks are listed in Table 1. 95% confidence intervals were estimated by bootstrapping the test set 1000 times.

The miss rate for a random allocation of antibiotics (preserving clinician budget constraints) was 21.5% (95% CI [18.6, 24.3]). The miss rate for the allocation made by clinicians was 15.7% (95% CI [13.3, 18.3]). The miss rate for the optimized antibiotic allocation (preserving clinician budget constraints) was 14.1% (95% CI [11.6, 16.5]).

By perturbing the budget parameters \( (K) \) in our problem formulation, we found that the optimized allocation yielded a miss rate equivalent to the actual clinician miss rate while reducing the Vancomycin & Piperacillin-Tazobactam budget in favor of a larger Piperacillin-Tazobactam budget by 69.8%. Similarly, an optimized allocation yielded a miss rate equivalent to the clinician miss rate while reducing the Piperacillin-Tazobactam budget in favor of Cefazolin by 40.2%, while reducing the Ceftriaxone budget in favor of Cefazolin by 33.4%, and while reducing the Ceftriaxone budget in favor of Ampicillin by 21.1%.

**Table 1:** Antibiotic Susceptibility Classifier Performance

<table>
<thead>
<tr>
<th>Antibiotic Selection</th>
<th>Best Model Class</th>
<th>Prevalence</th>
<th>Average Precision</th>
<th>AUROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vancomycin</td>
<td>Gradient Boosted Tree</td>
<td>0.23</td>
<td>0.46 [0.40, 0.52]</td>
<td>0.72 [0.68, 0.75]</td>
</tr>
<tr>
<td>Ampicillin</td>
<td>Gradient Boosted Tree</td>
<td>0.43</td>
<td>0.54 [0.49, 0.58]</td>
<td>0.62 [0.59, 0.65]</td>
</tr>
<tr>
<td>Cefazolin</td>
<td>Gradient Boosted Tree</td>
<td>0.59</td>
<td>0.72 [0.68, 0.76]</td>
<td>0.67 [0.64, 0.70]</td>
</tr>
<tr>
<td>Ciprofloxacin</td>
<td>Random Forest</td>
<td>0.63</td>
<td>0.73 [0.70, 0.76]</td>
<td>0.61 [0.58, 0.64]</td>
</tr>
<tr>
<td>Ceftriaxone</td>
<td>Gradient Boosted Tree</td>
<td>0.66</td>
<td>0.79 [0.77, 0.82]</td>
<td>0.69 [0.66, 0.72]</td>
</tr>
<tr>
<td>Cefepime</td>
<td>Random Forest</td>
<td>0.80</td>
<td>0.87 [0.84, 0.89]</td>
<td>0.65 [0.61, 0.69]</td>
</tr>
<tr>
<td>Vancomycin &amp; Ceftriaxone</td>
<td>Gradient Boosted Tree</td>
<td>0.81</td>
<td>0.87 [0.84, 0.89]</td>
<td>0.67 [0.63, 0.71]</td>
</tr>
<tr>
<td>Meropenem</td>
<td>Gradient Boosted Tree</td>
<td>0.82</td>
<td>0.90 [0.88, 0.92]</td>
<td>0.69 [0.65, 0.72]</td>
</tr>
<tr>
<td>Piperacillin-Tazobactam</td>
<td>Random Forest</td>
<td>0.90</td>
<td>0.94 [0.92, 0.95]</td>
<td>0.64 [0.59, 0.69]</td>
</tr>
<tr>
<td>Vancomycin &amp; Piperacillin-Tazobactam</td>
<td>Random Forest</td>
<td>0.96</td>
<td>0.98 [0.97, 0.99]</td>
<td>0.70 [0.62, 0.77]</td>
</tr>
<tr>
<td>Vancomycin &amp; Cefepime</td>
<td>Random Forest</td>
<td>0.97</td>
<td>0.98 [0.98, 0.99]</td>
<td>0.70 [0.62, 0.78]</td>
</tr>
<tr>
<td>Vancomycin &amp; Meropenem</td>
<td>Gradient Boosted Tree</td>
<td>0.98</td>
<td>0.99 [0.99, 0.99]</td>
<td>0.73 [0.65, 0.81]</td>
</tr>
</tbody>
</table>

**Discussion and Conclusions**

Tree based models consistently outperformed both l1 and l2 penalized logistic regressions in our twelve prediction tasks. The optimized antibiotic allocation performed significantly better than a random antibiotic allocation, but not significantly better than the clinician allocation. The optimized allocation yielded an equivalent miss rate to the clinician allocation while using fewer broad-spectrum antibiotics. More work is needed to evaluate how well these models would perform on the full deployment population (not just patients with positive cultures) and on prospective cohorts through use of the Epic Cognitive Compute platform which allows model predictions to display within medical records themselves. However, because of the benefits of reducing broad-spectrum antibiotic administration, this type of machine learning-based allocation may be advantageous compared to relying on clinician judgement.

**References**

Development and Validation of a Machine Learning Model to Predict Contact Moments in Post-Myocardial Infarction Home Measurements

Anne A.H. de Hond, MSc1, Esmee Stoop, MSc1, Nicole van Keulen, MSc1, Loes van Winden, MSc1, Ellen Poorter, MSc1, Douwe E. Atsma, MD, PhD1, Ilse Kant1, PhD, Ewout W. Steyerberg, PhD1

1Leiden University Medical Center, Leiden, the Netherlands

Introduction

Telemonitoring a patient after myocardial infarction (MI) could help detect clinical deterioration earlier which is associated with better clinical outcomes and higher patient satisfaction of care1-2. To this end, the Leiden University Medical Center provides its post-MI cardiovascular patients with ‘the Box’ consisting of several digital monitoring devices3. Patients are contacted by a nurse practitioner if their measurements indicate a significant clinical deterioration, referred to as a ‘contact moment’. However, the supervision of this daily incoming home measurements data is a time-consuming and taxing task for nurse practitioners. The resulting information overload can lead to burnout4 and poses a risk to patient safety5.

The aim of this study is threefold: i) develop a machine learning (ML) model that identifies high-risk patients in need of a contact moment, ii) validate the model in silent mode, and iii) identify areas for improvement in model performance and presentation based on the initial user experience of the nurse practitioners.

Methods

The data for this study concerns post-MI cardiovascular patients and was collected through the ‘the Box’ study3 at a tertiary care center in the Netherlands between 2015 and 2020. The contact moments were initiated by the nurse practitioner upon suspicion of a significant clinical deterioration based on the wearables data. Nurse practitioners later checked and adjusted the contact moments in the collected data to improve the quality of the outcome variable. Age, sex, weight, height, activity, blood pressure, heart rate and daily number of measurements were used in the modelling. The ratio of diastolic and systolic blood pressure and the body mass index (BMI) were calculated and added. For the blood pressure, heart rate and BMI, the mean and standard deviation (rolling window of 4 days) and the observations of the previous 6 days were included. A logistic regression, random forest and gradient boosted decision tree (XGBoost) algorithm were trained to predict relevant disease progression leading to unscheduled contact moments between patient and nurse practitioner, using 3-fold cross-validation. Performance was measured with the Area Under the Receiver Operating Characteristic Curve (AUROC) and Area Under the Precision-Recall Curve (AUPRC). Confidence intervals (CI) were calculated via bootstrapping. Shapley Additive exPlanations (SHAP) were added to estimate the importance of individual variables for the model predictions.

Figure 1. Time series for home measurements of patient with contact moment with a) systolic blood pressure, b) diastolic blood pressure, c) heart rate and d) BMI. Red line indicates a contact moment.
Upon completing model development, we will proceed by integrating the model in the IT infrastructure of the hospital and running it in silent mode, meaning that the predictions of the model will not be available to the users. In this pre-production environment, the model is prospectively validated on new incoming home measurements. Next, we will conduct user acceptance tests where nurse practitioners are able to interact with the model and provide feedback on its predictions and associated explanations.

Preliminary results

The data consisted of 373 MI patients telemonitored with ‘The Box’ for an average duration of 126 days. The patients were on average 59 years old and the majority was male (76%). Of these patients, 50 had one or more contact moments (13% of all included patients). Among all daily measurements, there were 88 contact moments (0.2% of total 46959 observations). For an indication of the progression for some of the clinically relevant variables see Figure 1.

The logistic regression model achieved an AUROC of 0.83 (0.73-0.92) and an AUPRC of 0.02 (CI 0.01-0.04, baseline 0.02). The random forest model obtained an AUROC of 0.86 (CI 0.77-0.94) and an AUPRC of 0.03 (CI 0.01-0.07, baseline 0.02). The XGBoost model achieved an AUROC of 0.86 (CI 0.77-0.94) and an AUPRC of 0.03 (CI 0.01-0.12, baseline of 0.02). The random forest and XGBoost models performed similarly and slightly better compared to the logistic regression. The XGBoost was selected to be implemented in clinical practice, because of its ability to easily process missing data and good processing speed. The strongest predictors based on SHAP values were those related to systolic blood pressure and daily number of measurements.

We expect good prognostic validation results, as the intended population is the same for development and validation. The nurse practitioners are likely to have some very valuable input on clinically meaningful thresholds for probability predictions, acceptable levels of false positives and false negatives, and presentation of results and explanations.

Conclusion

We developed a model that can predict the need for unscheduled consultation between ‘The Box’ patients and nurse practitioners to address disease progression. An accurate prediction of contact moments between patient and nurse practitioner aids in prioritizing the incoming home measurements. The nurse practitioners can then immediately direct their attention to the most urgent cases. This has the potential to save time and energy and can address the problems related to information overload on the clinician’s side, like burnout.

Clinician involvement is crucial to promote a successful implementation of new technologies. We therefore closely work with the nurse practitioners during development and validation of the algorithm. Their input will be used to improve the algorithm and its presentation.

References

Using Genomic Association Replication Rates as an EHR Quality Measure via the Phenotype-Genotype Reference Map (PGRM)

Sarah DeLozier, BS,1 Jing He, MS,1 Josh Peterson, MD, MPH, FACMI,1 Josh C. Denny, MD, MS, FACMI,2 Lisa Bastarache, MS1
1Department of Biomedical Informatics, Vanderbilt University, Nashville, TN
2All of Us Research Program, National Institutes of Health, Bethesda, MD

Introduction: Genome-wide association studies (GWAS) of large cohorts have greatly enhanced our understanding of the pathophysiology of complex disease. Secondary use of clinical data from the electronic health record data (EHR) would benefit from validated quality control methods to ensure the reliability of results.1 We created a phenotype-genotype reference map (PGRM) consisting of single nucleotide polymorphism (SNP)-trait associations from the GWAS catalog (SNP/phenotype pairs). We mapped phenotypes from the catalog to PheWAS codes (phecodes) and used reported effect sizes and allele frequencies to calculate the number of cases required for replication at 80% power. The PGRM map contains all associations in the GWAS catalog which have an exact matching phecode, a total of 9,770 SNP/phenotype pairs, and is annotated with the genetic ancestry of subjects from the original study. The PGRM can be used prior to GWAS to calculate a replicate rate for EHR cohorts that are linked to genetic data. We propose using the PGRM as an EHR data quality measure for GWAS and as a method of testing the effect of standard GWAS parameters (e.g., minimum code count) on replication.

Methods: Developing the phenotype/genotype reference map reference map (PGRM)

Figure 1 illustrates the workflow of the phenotype/genotype reference map (PGRM). Phenotypes in the map are represented by phecodes, groups of International Classification of Disease-9, 10, and 10-CM billing codes developed to support phenome-wide association studies (PheWAS).3 The GWAS catalog diseases and traits are annotated with the Experimental Factor Ontology (EFO). We mapped EFO terms used in the catalog to phecodes through manual review and annotated those that were exact matches. Inexact matches between EFO/phenocodes were characterized by type of match and those overly broad matches (e.g., phecode 305.2 "Eating disorder" mapped to the EFO "Bulimia nervosa") or narrow matches (e.g., phecode 627.4 "Premenopausal menorrhagia" mapped to EFO "Menorrhagia") were excluded. Continuous traits and associations not reaching genome wide significance (i.e., p-value <5x10^-8) were also excluded. Lastly, we annotated associations by study ancestry, derived from the ‘Initial Sample’ and ‘Replication Sample’ fields in the GWAS catalog. Information missing from the GWAS catalog was ascertained manually from the original manuscript. Power calculations were conducted using the average effect size reported in the catalog and ancestry matched allele frequencies reported in GnomAD.

Calculating replication rate

We attempted to replicate all powered associations in the PGRM in a local DNA biorepository (BioVU) cohort of 65,215 individuals of European ancestry, age >=18, genotyped on the MEGA array. SNPs with minor allele frequency of <1% and those with missingness > 5% were excluded. Cases were defined as having at least two phecodes, the default minimum code count (MCC) threshold used in PheWAS studies.3 Associations were included in the analysis when the number of cases in the BioVU cohort were equal to or exceed the “cases needed” in the PGRM to achieve 80% power. The replication rate for the cohort was calculated by dividing the number of SNP/phenotype pairs with p<0.05 by the total number of SNP/phenotype pairs attempted in replication. In total, we attempted replication on 1,428 SNP/phenotype pairs, and 95 unique phecodes across 12 different disease categories.

Comparing replication rates

We used the PGRM to compare the replication rate of associations using different minimum code counts (MCC) in the BioVU cohort. We tested MCC 1-4 and calculated the replication rate and the number of associations powered in each analysis (Figure 2). Replication rates between MCC were compared using a one-sample test of proportions.

Results: Overall, the PGRM includes 9,770 SNP/phenotype pairs for 173 unique phecodes. By ancestry, the European genetic ancestry map is the largest, with 7,981 unique SNP/phenotype pairs, compared with 295 and 1,686 for the African ancestry and Asian ancestry maps respectively. The density of the PGRM differed by phecode category, with neoplasms having the highest number of associations (1,779), followed by mental disorders (1,343). Congenital anomalies had the fewest associations (24), followed by infectious disease (119).

Using the BioVU cohort, we attempted to replicate 1,482 associations in European ancestry PGRM. Comparing different MCC thresholds, we found a significant improvement in replication rate between MCC=1 and MCC=2.
(59% to 67%, p=8x10^{-6}). There was no significant difference in replication rates above a MCC of 2. Increasing MCC also resulted in a decrease in SNPs that were adequately powered for replication (Table 1).

**Discussion**

Meta-analyses of GWAS findings suggest high external replication rates of SNP/phenotype associations.\(^4\) Researchers using EHR-based cohorts can similarly instill confidence in their results through replication. Low replication rates in a given dataset may help identify incorrect or implausible data not fit for GWAS.\(^1\) The PGRM enables the calculation of replication rates for novel datasets and continued opportunity to validate known GWAS associations, including testing associations across ancestries. Thus, through known associations in the GWAS catalog, the PGRM encourages replication as a part of the QC process.

While the PGRM does include a diverse set of 173 phenotypes, not all diseases are equally represented in the map. For example, as in the GWAS catalog, infectious diseases are not well replicated in the PGRM. Furthermore, the European ancestry map is far more robust than that of Asian and African ancestries, reflecting a well described ancestry-based disparity in genetics research.\(^5\) The PGRM should be used to optimize GWAS models, including but not limited to minimum code count criteria, that are likely phenotype-specific.

**Key Conclusions**

In this study, we modeled common data problems and tested GWAS methods predictive of replicating known associations from the GWAS catalog, finding that adding minimum code counts increased replicating total possible associations replicated while decreasing power. Future work should assess the assimilation of the PGRM in GWAS and its impact on the ability of a novel dataset from the health record to predict known or novel associations.

**Table 1:** Replication rates for increasing minimum code counts (MCC).

<table>
<thead>
<tr>
<th>Analysis</th>
<th>Powered SNP/phenocode associations</th>
<th>Replicated associations</th>
<th>Percent Replicated</th>
</tr>
</thead>
<tbody>
<tr>
<td>MCC=1</td>
<td>1,842</td>
<td>1,092</td>
<td>59.3%</td>
</tr>
<tr>
<td>MCC=2</td>
<td>1,428</td>
<td>956</td>
<td>66.9%</td>
</tr>
<tr>
<td>MCC=3</td>
<td>1,259</td>
<td>861</td>
<td>68.4%</td>
</tr>
<tr>
<td>MCC=4</td>
<td>1,149</td>
<td>791</td>
<td>68.8%</td>
</tr>
</tbody>
</table>

**Figure 2:** Percent replication by phenotype (phecode) disease category using a cohort from a local DNA biorepository, BioVU (minimum case count =2). The number of SNP/phenotype pairs attempted to replicate is shown adjacent to the section label.

**References**


**Funding Source:** This work was supported by P50GM115305 from NIH/NIGMS.
Can reproducibility be improved in clinical natural language processing?  
A study of 7 clinical NLP suites

William Digan, PhD,1,2, Aurélie Névésol, PhD,3, Antoine Neuraz, MD, PhD,1,4, Maxime Wack, MD,1,2, David Baudoin,1,2, Anita Burgun, MD, PhD,1,2, Bastien Rance, PhD1,2

1INSERM, Centre de Recherche des Cordeliers, UMRS 1138, Université de Paris, Université Sorbonne Paris Cité, Paris, France; 2Department of Medical Informatics, Hôpital Européen Georges Pompidou, Assistance Publique–Hôpitaux de Paris, Paris, France; 3Université Paris Saclay, CNRS, LISN (formerly, LIMSI), Orsay, France; 4Department of Medical Informatics, Necker Children’s Hospital, Assistance Publique–Hôpitaux de Paris, Paris, France

Introduction
The increasing complexity of data streams and computational processes in modern clinical health information systems makes reproducibility challenging. Clinical natural language processing (NLP) pipelines are routinely leveraged for the secondary use of data. However, the practical implementation of state-of-the-art methods that have been validated in NLP research studies proves challenging in a clinical operational environment. Translational research would benefit from facilitating the transition of experimental NLP methods to clinical NLP tools. This is especially true for languages other than English, for which a variety of tools are developed outside of large NLP pipelines. Faced with similar challenges, the field of bioinformatics has developed over the years multiple solutions to help tackle these challenges. In particular, workflow management systems (WMS) have been widely used to handle the reproducibility bottleneck. In a recent study1 we aimed to evaluate whether WMS and other bioinformatics practices could impact the reproducibility of clinical NLP frameworks.

Materials and Methods
Based on the literature across multiple research fields (NLP, bioinformatics and clinical informatics) we selected articles which (1) review reproducibility practices and (2) highlight a set of rules or guidelines to ensure tool or pipeline reproducibility. We aggregate insight from the literature to define reproducibility recommendations. Finally, we assess the compliance of 7 NLP frameworks to the recommendations.

Results
We selected 8 selected articles5,6,7 from 455 articles discussing reproducibility retrieved from Web of Science and PubMed. We identified 5 topics of reproducibility (namely, traceability, versioning, standardization, usability, and shareability) covering 40 recommendations. Frameworks based on WMS match more than 50% of features (26 features for LAPPs Grid,8, 22 features for OpenMinted) compared to 18 features for current clinical NLP framework (cTakes,9, CLAMP10) and 17 features for GATE11, ScispaCy12, and Textflows13. This indicates that the community is well aware of the need for reproducibility and already making valuable efforts towards compliance. While current systems integrate many criteria in spite of structural limitations, there is still room for growth.

Discussion
All recommendations are largely described by NLP and bioinformatics communities. 34 recommendations are endorsed by at least 2 articles from our selection. Overall, 15 features were adopted by every NLP Framework. Nevertheless, frameworks based on WMS had a better compliance with the features. This suggests that the clinical NLP community can benefit from integrating reproducibility guidelines that have been formulated and sometimes implemented in various communities. In particular, WMS combined with modular design and integration of tools (regardless of programming languages and specifications) shows great potential at limited cost. In addition, documenting provenance of components and data can ensure the traceability and usability of analysis. Finally, curated lists of annotated datasets, programming libraries, containerized tools and pipelines could facilitate the re-usability of shareability of research results.
Conclusion
In summary, we identified 40 reproducibility recommendations based on the review of eight articles from heterogeneous research fields (bioinformatics, medical informatics and natural language processing) and our daily experience of deployment and use of NLP processes in a clinical setting. NLP frameworks can take advantage of lessons learned from other fields (and especially bioinformatics) to improve reproducibility for NLP systems in clinical settings. More precisely, specific features could be transferred to clinical NLP such as public repositories of curated workflows, enhanced modularity and shareability with containers or provenance information for traceability of processes and results. We believe that reproducibility is a necessary - although not sufficient - intermediate step towards the reuse of NLP tools, including modern neural methods. Versioning and reproducibility (including distribution of sample open data) are actionable steps that augment FAIR principles to empirically verify the validity of tools used in a new environment.

References
Alerting Primary Care Teams to Out-of-Network Acute Care Events using Health Information Exchange: Impact on Timely Follow-Up

Brian E. Dixon, MPA, PhD1,2,3; Kimberly M. Judon, MPH4; Ashley L. Schwartzkopf, MSW1; Vivian M. Guerrero, MA4; Nicholas Koufacos, MSW4; Justine May, MSW1; Cathy C. Schubert, MD1; Kenneth S. Boockvar, MD, MS4,5

1Department of Veterans Affairs, Center for Health Information and Communication, Indianapolis, IN; 2Indiana University Fairbanks School of Public Health, Department of Epidemiology, Indianapolis, IN; 3Regenstrief Institute, Center for Biomedical Informatics, Indianapolis, IN; 4Department of Veterans Affairs, James J. Peters VA Medical Center, Bronx, NY; 5Icahn School of Medicine at Mount Sinai, New York, NY

Introduction
Providers seek to leverage health information technologies to access information on their patients, regardless of its source. Health information exchange (HIE) is the electronic transfer of clinical, administrative, or other information necessary for the delivery of health care across diverse systems or organizations. Most often HIE takes the form of an information system transaction mediated by technical standards, such as the electronic reporting (or pushing) of a result from a laboratory information management system to an electronic health record (EHR) system using the Health Level Seven (HL7) messaging standard. Other times HIE involves querying (or pulling) information from another system, such as the transmission of a summary of care record using the HL7 Consolidated Document Architecture (CDA) standard outlined in meaningful use criteria for Medicare and Medicaid providers (1).

Event notification is a form of HIE involving the electronic reporting (or pushing) of information pertaining to a clinical event from one provider to another facilitated by a messaging standard. Notification usually pertains to acute care events (e.g., hospitalization, emergency care), and the notifications are typically sent to primary care providers responsible for coordination of care (2). To date there has been just one quantitative study of event notifications sent to ambulatory providers for acute care events over a 3.5 year period using a regional HIE network (3). The study found a statistically significant 2.9% reduction in 30-day readmissions observed during the intervention period.

Our objective was to examine the effectiveness of event notification on health care delivery processes and outcomes for patients with out-of-network events. Given the ability of HIE to facilitate access to patient information, especially after a handoff or transfer of care, we hypothesized that HIE would impact care coordination activities, including timely follow-up and reintegration into primary care following an acute care episode.

Methods
We employed a concurrent cohort study involving two groups: 1) patients whose primary care team received an alert following an acute care event at an out-of-network facility; and 2) patients for whom no alerts were sent (e.g., controls). All patients were older Veterans (age ≥65 years) who received primary care at the following U.S. Veterans Administration (VA) medical centers: 1) James J. Peters VA Medical Center (JJP VAMC) located in the Bronx, New York; and 2) Richard L. Roudebush VA Medical Center (RLR VAMC) located in Indianapolis, Indiana.

This study is part of a larger, cluster randomized controlled trial at the two VA medical Centers. The protocol for the larger study is published on ClinicalTrials.org (NCT02689076) and available online (4). The study was reviewed and approved by the Institutional Review Board (IRB) of Indiana University as well as the VA Research & Development Committee at both the Indianapolis VA Medical Center and the Bronx VA Medical Center.

Upon utilization of non-VA acute care, an HL7 admission-discharge-transfer (ADT) message was electronically sent from the non-VA acute care facility to an HIE network connected to the two participating VA sites. Each HIE network responded to the ADT message by notifying (or alerting) the VA that an enrolled patient visited a non-VA care facility. Following notification, study coordinators at the two VA sites sent an internal electronic note within the VA’s EHR system to the Veteran’s primary care medical home team, referred within the VA as a Patient-Aligned Care Team (PACT). This note, which becomes part of the Veteran’s medical record, identifies the non-VA care facility and provides information on the reason for the visit or chief complaint. Each note required acknowledgement of receipt by a PACT team member. Alerts were sent between March 2016 to December 2019.
This analysis examined each eligible non-VA encounter during the study period. We examined the notes entered by PACT team members to ascertain whether the Veteran was contacted within 7 days following the non-VA discharge and/or whether the Veteran had a follow-up appointment within 30 days following non-VA discharge.

Baseline characteristics of patients were assessed using descriptive statistics. Comparisons in baseline characteristics between the two arms were examined using \( \chi^2 \) test or Fisher exact test for categorical variables and t-test for continuous variables. Since the units of randomization were primary care teams rather than Veterans, there was potential for imbalance between groups in terms of baseline covariates due to clustering. Group comparisons were analyzed using PROC GLIMMIX for outcomes assuming a random intercept and binary distribution while accounting for clustering within primary care teams. P-values and odds ratios were calculated for comparisons between groups, adjusting for baseline group differences in age, race/ethnicity, and Medicaid status.

**Results**

Overall the mean age of participants was 74.8 years, and the cohort was almost exclusively male (98.0%) which is typical in VA-based studies. Enrollment site and the type of non-VA acute care received during the study period was balanced, with no differences between arms. Age did vary by arm, with the usual care group mean age younger (72.5 years) than the notification group (76.9 years, \( p<0.01 \)). More non-Hispanic black Veterans were in the usual care group (30.4%) compared to the notification group (17.1%, \( p<0.05 \)).

Compared to usual care, Veterans whose primary care team received notification of their non-VA acute care encounter were 4-times more likely to have phone contact within 7 days (AOR=4.10, \( p<0.001 \)) and 2-times more likely to have an in-person visit within 30 days (AOR=1.98, \( p=0.007 \)) with a primary care provider (PCP, medical doctor or nurse practitioner) after discharge. When follow-up was expanded to include registered nurses on the primary care team, Veterans whose team received acute care notifications were almost 3-times more likely to have phone contact or in-person visit (AOR=2.65, \( p<0.001 \)) within 30 days of the acute care event.

**Discussion**

There exists a strong theoretical case for HIE to impact care coordination activities, such as reintegration into primary care following an acute care episode. We found that when primary care teams are notified of out-of-network acute care episodes, they respond by initiating contact with the patient via phone or by scheduling an in-person visit. In pre-implementation conversations with primary teams, clinicians said they often learn about non-VA encounters weeks or months after the acute care encounter, if they learn about them at all. Our findings provide evidence that care teams will use event notifications to coordinate follow-up care. Yet, more research is necessary to examine the impact of event notification on patient and population outcomes. Moreover, additional informatics research will be necessary to optimize integration of event notifications into workflow across the VA and outside the VA health system. Not all providers acted upon the alerts, even if receipt was acknowledged. Healthcare workers beyond providers might play critical roles in the process. There may be other technologies that could facilitate the process, including patient portals and decision support that could prioritize patients for follow-up care based on encounter diagnostic codes. While some non-VA events may have been missed by the system, each regional HIE network covered >90% of the geographic areas and populations served by the two VA medical centers minimizing potential for missed events.

**Acknowledgements**

The project described is supported by Merit Review Award Number I01 HX001563 from the VA Health Services Research & Development Service of the VA Office of Research and Development. The protocol and opinions are solely the responsibility of the authors and do not necessarily represent the official views of the VA.

**References**

Informing Symptom Science Using a Citizen Science Application in the COVID-19 Pandemic

Caitlin Dreisbach, PhD, RN¹, Katherine South, BSN, RN, CPN², Theresa Koleck, PhD, RN³, Veronica Barcelona, PhD, MSN, MPH, RN, PHNA-BC², Lena Mamykina, PhD⁴, Noemie Elhadad, PhD, FACMI⁴, Suzanne Bakken, PhD, RN, FAAN, FACMI²

¹Data Science Institute, Columbia University, New York, NY, USA; ²School of Nursing, Columbia University, New York, NY, USA; ³School of Nursing, University of Pittsburgh, Pittsburgh, PA, USA; ⁴Department of Biomedical Informatics, Columbia University, New York, NY, USA

Abstract

In April 2020, researchers at Columbia University deployed COVIDWATCHER, a citizen-science platform to understand the impacts of the novel coronavirus pandemic. This mixed-methods study examined the self-reported symptoms from COVIDWATCHER to compare the symptom experiences of users who identify as female compared to male, and further, for users who identify as Hispanic compared to non-Hispanic. Our results show a difference between reported symptoms that are likely attributed to social and environmental factors.

Introduction

In response to the surge in the novel Coronavirus disease (COVID-19), a collaborative team at Columbia University built COVIDWATCHER, an application to help inform the response to COVID-19 and better understand the personal, social, and health implications of the pandemic. COVIDWATCHER is driven by the approach of citizen science, or the public actively participating in scientific research.

Medical providers during the pandemic have often needed to make a COVID-19 diagnosis based only on symptoms. Major health organizations have already established the differential impact of the COVID-19 pandemic on women and girls [1] and, further, on women of color, including Hispanic women [2] and African American women [3]. Hispanic women across their life course have already been noted to be a demographic group at high risk for testing positive for COVID-19.

Even before the COVID-19 pandemic, women have been reported to share the same medical diagnosis differently than men [4]. The most famous example of this is the somatic symptoms during a heart attack. Though both men and women experience severe chest pain, women tend to present to emergency rooms with nausea, increasing fatigue, and shortness of breath compared to men who predominantly focus on the chest discomfort [5]. The core concern is that much of the research underlying the known biology has historically been conducted in men has contributed to gender bias in research [6].

This mixed-methods study utilized the data from 551 (49.6%) users who report their gender as female. Across the entire COVIDWATCHER user sample who answered the symptoms attestation, 10.4% of the sample registered their ethnicity as Hispanic.

We hypothesized that there were symptom phenotypes based on health status, environmental factors, and access to healthcare that impact a participant’s health trajectory throughout the pandemic and that these phenotypes are different for users who identify as female compared to

| Number of users who answered the symptoms survey | 1,109 |
| Number of user encounters | 15,455 |
| Gender | N (%) |
| Male | 529 (47.8%) |
| Female | 551 (49.6%) |
| Nonbinary | 13 (1.2%) |
| Transgender | 1 (<0.1%) |
| Other* | 15 (1.3%) |
| Hispanic heritage | N (%) |
| Yes | 115 (10.4%) |
| No | 972 (87.6%) |

Table 1. Participant characteristics in COVIDWATCHER.
those who identify as male users. Further, we hypothesized that there would be different symptom phenotypes between Hispanic and non-Hispanic women.

**Combined Qualitative and Quantitative Analysis of Reported Symptoms**

Our team completed a clustering analysis of reported symptoms from participants in COVIDWATCHER. As shown in Figure 1, there was a visual difference in the reported clustering of participants from users who identify as female and Hispanic/Latina (n=64) compared to non-Hispanic females (n=481) and users who identify as male of all ethnicities (n=529).

Labeling of the clusters based on the rich qualitative interviews found that clusters represented diagnosis variation (i.e., healthy, COVID-19, vaccine side effects, allergies, or stress). Clusters that represented COVID-19 symptomology were further characterized to show highly predictive symptoms such as fever, loss of taste, and loss of smell as driving symptoms in that cluster. Themes in the qualitative interviews expanded on both social and environmental stressors, particularly for women of color, that contributed to symptom development.

**Conclusion**

The mixed-methods results from this study highlight a discordance between when a participant uses COVIDWATCHER and when they experienced COVID or COVID-like symptoms. Clustering using COVIDWATCHER data may not be able to discern between healthy or seasonal allergies diagnoses but may be able to identify COVID for a smaller group of participants. Limitations of self-reported symptoms and a lack of training data with confirmed COVID symptoms make it difficult to answer our original hypothesis. More rigorous confirmed symptomology would be needed to fully test the hypothesis that women experience COVID disease differently than men. However, use of citizen science platforms during the pandemic can yield important insight into both health and social impacts of this unique time period.

**Acknowledgements**

This research is funded as a pilot study through the Precision in Symptom Self-Management (PriSSM) Center (P30 NR016587).

**References**

Clinical Decision Support in the Pediatric ICU: A Multi-Institution Survey

Adam C. Dziorny, MD, PhD1; Julia A. Heneghan, MD2; Moodakare Ashwini Bhat, MD3; Dean J. Karavite, MSI4; L. Nelson Sanchez-Pinto, MD, MBI5; Jennifer McArthur, DO6; Naveen Muthu, MD4

1University of Rochester School of Medicine, Rochester, NY; 2University of Minnesota Masonic Children’s Hospital, Minneapolis, MN; 3Medical College of Wisconsin, Milwaukee, WI; 4Children’s Hospital of Philadelphia, Philadelphia, PA; 5Northwestern University Feinberg School of Medicine, Chicago, IL; 6St. Jude Children’s Research Hospital, Memphis, TN

Introduction: Widespread adoption of electronic health records (EHRs) has led to significant growth in the volume of information available to clinicians in real time. The pediatric intensive care unit (PICU) is a data-intense environment, with hundreds of thousands of data points generated daily, including laboratory results, streaming vital signs, and ventilator parameters. Clinical decision support (CDS) systems are primed to leverage these rich data, however, most of the CDS in use are rule-based alerts, order sets, and order entry systems. While many studies have focused on the development of novel predictive models, real-world implementations using CDS have lagged. Broad consensus around the goals and challenges of CDS in the PICU is lacking. In reviewing CDS in the PICU over a decade ago, Mack et al. concluded that while effectiveness was variable, factors such as incorporating into workflows, enhancing usability and considering human factors were essential to achieving success.1 This remains true today. However, to fully realize the potential for CDS in the PICU, we need to address critical gaps including understanding CDS usage, utility, and usability.

Objective: Describe the current state of PICU CDS, perceived goals and challenges, and evaluation approaches.

Methods: This IRB-approved multi-institution survey followed a published seven step development process.2 Following a literature review, semi-structured interviews were completed and coded by 3 authors to identify relevant themes. From these, survey themes were developed, and subsequent items were iteratively developed and grouped into themes. Surveys were implemented in REDCap; evaluation and guided exploration were performed at multiple sites. Finally, surveys were piloted at author institutions and updated before distribution.

<table>
<thead>
<tr>
<th>Table 1. Demographics of survey respondents by group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medical Director (N=27)1</td>
</tr>
<tr>
<td>--------------------------</td>
</tr>
<tr>
<td>EHR in ICU 1 (%)</td>
</tr>
<tr>
<td>Epic Systems Cerner</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Unit Type 2 (%)</td>
</tr>
<tr>
<td>Med / Surg Cardiac</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td># Beds, Median (IQR)</td>
</tr>
<tr>
<td>Years since last training, Median (IQR)</td>
</tr>
</tbody>
</table>

1Remaining indicated a mixture of other EHRs in use. 2Remaining indicated mixed units.

Free text responses were coded independently and discrepancies were resolved by discussion. Descriptive statistics are presented as n (%) or median (IQR) unless noted. Data analysis was performed in R Studio.

Results: Survey was organized into themes identified from 10 semi-structured interviews included current CDS usage, perceived limitations of CDS use, goals and perceived concerns of CDS use, and CDS evaluation. A total of 109 surveys were completed from 45 institutions. Demographics and counts per respondent group can be found in Table
1. Groups were similar in unit type and bed numbers; medical directors were further from last training and attendings were less likely to use Epic Systems EHR compared to Cerner.

**Current CDS Usage:** CDS were divided into EHR and non-EHR tools. In all groups, most-used EHR tools included passive CDS (“Lab result highlighting”, “Abnormal vital signs”) as well as active CDS (“order sets”). Least used were “Predictive tools” and “Non-order-based alerts.” Most-used non-EHR tools included “Bedside monitors” and “People-based resources”, while least-used included “Dashboards”, “Medical calculators” and “Predictive algorithms.” Custom EHR predictive tools (38%) included sepsis alerts (N=25, 61%), acuity scores (N=14, 35%), COVID-related (N=1, 2%) and other (N=11, 27%). Some responses included multiple categories and thus percentages do not add to 100%.

**Perceived Limitations:** All groups agreed that “use of order sets saves time” and “abnormal lab highlighting is useful”; most respondents were neutral or disagreed that EHR and non-EHR “predictive algorithms provide useful information” (Figure 1). **Goals and Concerns:** CDS goals common to all respondent groups include “evidence-based” CDS, “well-placed and delivered at the right time”, with a “proven impact on patient safety”. Common concerns include “unknown decision support accuracy”, and practitioners “losing the ability to think independently”. **CDS Evaluation:** Among all groups, feedback on CDS use and burden is rarely or never provided, workflow observations are not traditionally part of CDS development, and systems to understand alert burden are non-existent.

**Discussion:** In this multi-institution survey, we identify current practice patterns and perceptions of CDS use across multiple provider roles in the PICU. Interestingly, we identified both in survey development and in the survey results that clinicians’ description of CDS broadly includes resources not traditionally considered “CDS” by informaticians, such as people-based resources (pharmacists, dieticians). Passive CDS was generally considered more useful than active CDS. Perhaps because of the slow integration of predictive algorithms with clinical systems, perception of predictive algorithms was poor across all groups. These usage patterns and perceptions identify areas of focus for targeting CDS to particular groups, as well as future exploration of under-developed CDS space such as translating evidence-based predictive models into clinical care.

The identified goals and concerns of this broad group of practitioners can be used to establish a framework to evaluate CDS in the PICU. Following development, evaluation of CDS use and performance is critically under-developed across all respondents. While most felt that feedback on CDS was well-received, use and alert burden was rarely tracked and provided to users. This space provides ample opportunity for research and operational support.

Overall, practitioners have shared goals and concerns regarding CDS and its implementation. It should be a priority for institutions and the larger ICU informatics community to support thoughtful development and implementation supporting these priorities.

**References**

Development and Global Use of a Platform-Independent Mobile App to Enable Citizen-Scientist Data Collection on Mask-Wearing

Peter J. Embi, MD, MS1, Marcelo A. Lopetegui, MD, MS2, Joshua R. Vest, PhD1
1Regenstrief Institute and Indiana University, Indianapolis, Indiana, USA; 2ICIM-Universidad del Desarrollo, Santiago, Chile.

Introduction: Wearing face masks in public is one of the easiest and least expensive strategies to control the COVID-19 pandemic. When properly worn, masks can effectively filter out respiratory droplets that may contain SARS-CoV-2. Policies mandating mask wearing have been associated with reduced risk of community-wide transmission of COVID-19. Despite this, mask wearing has been inconsistently adopted by the public, and self-reported mask behavior has varied over time, geography, and demographic groups. Further, data on proper masking wearing and from direct observational studies of mask wearing are limited.

Approaches that enable public data collection and aggregation via mobile apps have proven successful for crowd-sourcing activities such as monitoring traffic patterns. Mobile apps have also been used to enable scientific data collection by the public in fields ranging from ornithology and entomology to biomedicine. During the COVID-19 pandemic data on mask-wearing behavior was scant, but sorely needed. Potential benefits to direct observational data collection included correlating mask-wearing data with other public health data on COVID-19 disease activity, assessing the impacts of public health policies and mandates, and providing real-world data on mask usage to inform public health policies and educational campaigns.

To enable crowd-sourced data collection of directly observed public mask-wearing behavior, whether by the general public (citizen scientists) or even by trained public health workers and field researchers, we developed a Web-based mobile app. Importantly, we designed the app to be used discretely and to ensure anonymity of those being observed, while still enabling the easy and accurate capture of observed mask-wearing by the user.

Methods: Given the time-sensitive nature of the pandemic and the need to quickly create and maintain a cross-platform mobile app that could work across the globe in a low-cost manner, we elected to use Web-based technologies for the rapid development of our mobile app. User centered design was used, incorporating early adopters’ feedback on every sprint. Iteratively, the user interface design team created an extremely simple means for capturing mask wearing observations.

The app was built on a serverless architecture-based technology, Google Firebase, chosen for its flexibility, low costs, and given the world-wide scalability required. Internationalization of the user interface was achieved by using automated, artificial intelligence-based translation of text from the original English to 15 total languages at the time of release. Several languages considered the most likely to be used were manually validated by native speakers, including German, Italian, Portuguese and Spanish. Upon account creation, users were informed that the application would collect location data only while the app was in use, that their information would remain confidential, and they were asked to accept a user agreement that included an informed consent to use their deidentified data for secondary analyses. Each observation was represented by a dot on a map on the users’ screen that reflected their precise location. If needed to improve data accuracy, users could correct data (e.g. move the location of an observation, change the type of observation, increase or decrease the count observed, or delete the observation). This project and the use of the tool and data for research was reviewed and approved by the Indiana University Institutional Review Board.

Results: In under 3 weeks, a beta-version of the software was deployed for initial testing. Feedback from testers informed each development sprint, leading to the first stable version and release on September 12, 2020, just 5 weeks after conceptualization of the app. Based on user feedback, several enhancements were made during the Beta-testing and early release periods. Given the need for discrete data capture, often while the user was moving through a public space, we evolved the interface from a simple button-based data collection approach for noting “mask” “partial mask” or “no mask” to one that also included the well-known swipe gesture, commonly used for dating apps (Figure 1). They achieved this by swiping to the left, down, or right, respectively. Beta users and early adopters reported this was easy to use the app ‘on the move,’ while maintaining attention on their surroundings and not making it obvious they were collecting data. Over time, users requested, and we built-in, more registration and login options, including multiple social media authentication services, email and password, and even password-less login via SMS or email links. To encourage repeated use of the app, gamification techniques were also enhanced over time, such as the earning of

1356
badges with increasing levels of ‘prestige’ the more sessions a user completed. Registered users could also view more detailed, aggregate data on mask-behavior around the vicinity of the user could also be viewed on the app itself upon registration for an account (Figure 1).

At the time of public release, we also created a Website (www.maskcount.com) to provide easy access to the app, detailed information about the app, including its purpose, and frequently asked questions. To encourage adoption and demonstrate public value, a separate dashboard was subsequently added to the Website where anyone could visualize aggregate data generated by Maskcount users in real-time (Figure 2).

From release through July 27, 2021, 1321 individuals across 13 countries registered accounts. Among those, 511 actively used the app to collect 93,892 observations collected across 3339 sessions. In addition to use by the public, public health practitioners, researchers and educators reported use of the app and have started to publish findings based on its use.

Discussion: The rapid creation and deployment of a Web-based app designed to enable anonymous data collection by ‘citizen-scientist’ resulted in substantial adoption and use by interested individuals across the globe, and it provided new data that was used by some to gain new public health insights.

Despite the global pandemic, even greater widespread ‘viral’ adoption was not observed. Targeted use of such apps by public health, academic and research professionals was notable and could have implications for similar approaches. Further details about the platform and the data generated, a live demo, and additional lessons learned will be discussed during the presentation.

Reference

Advancing National HIE Measurement by Using Shared Patients Volume

Jordan Everson, PhD\textsuperscript{1} and Julia Adler-Milstein, PhD\textsuperscript{2}
\textsuperscript{1}Office of the National Coordinator for Health IT at HHS; \textsuperscript{2}University of California, San Francisco, California, USA

Introduction
Advancing health information exchange (HIE) has been a top policy priority for more than a decade. National measures of HIE adoption have been reported as the percent of hospitals and physicians that respond on surveys that they engage in HIE; specifically, survey questions ask if they electronically find/query, send, receive, and integrate data from outside sources (e.g., from other hospitals, health systems).\textsuperscript{1} However, these measures fail to capture heterogeneity in HIE connectivity – a given provider may be able to exchange with some providers but not others based on the HIE network(s) in which they participate. Importantly, these measures also do not capture the extent to which information is shared between organizations that commonly share in the treatment of patients. Providers may not be able to share information with other organizations that participate in HIE because HIE networks have developed unevenly across the country, with hundreds of different networks that offer varied approaches to connectivity – some focus on a given region, others focus on customers of a given EHR vendor, and still others focus on a care setting (e.g., emergency departments). As a result, provider organizations face a complex array of options and must decide in which network(s) to participate.\textsuperscript{2}

These networks are not necessarily optimally distributed, such that hospitals are able to engage in HIE with the other hospitals that treat the same patient population. For example, Epic Systems has created a network that allows hospitals on Epic to engage in HIE with other hospitals on Epic (vendor-centric HIE networks) but many care transitions occur between Epic and non-Epic provider organizations. Hospitals also choose to join HIE networks that are vendor-agnostic and have been established in some (but not all) regions. However, these HIE networks often serve a defined geographic area that may not match where patients travel for their healthcare. In addition, most states have multiple vendor-agnostic networks in overlapping geographic regions and, in recent national data, only 47% of eligible hospitals (i.e., those in the targeted region) decided to join any network.\textsuperscript{3} We therefore sought to create improved, patient-centered measures of national HIE connectivity based on the proportion of shared patients for which treating hospitals are able to exchange information because they connect to the same network.

Methods
We combined three datasets to generate measures for 2018. The first was a publicly-available dataset containing pairs of hospitals and the annual volume of shared fee-for-service Medicare patients (defined by both providing treatment to the same patients within the calendar year).\textsuperscript{4} We combined this pair-level data with information from Definitive Healthcare (formerly HIMSS) and from the American Hospital Association IT Supplement Survey on the specific HIE networks (including both health information organizations (HIOs) and electronic health record (EHR) vendor networks) that hospitals participated in each year. Using these data, we measured the proportion of patients shared between pairs of hospitals that were connected by HIE in 2018. Finally, we created multivariable regression models examining characteristics of the hospital pairs associated with participation in the same HIE network. Our analyses include all non-federal, acute-care hospitals in the United States included in all three datasets (1,721), comprising 16,433 hospital pairs that had a total of 6,492,232 shared patients in 2018. Analyses used sampling weights to adjust for differences in characteristics between hospitals in our analytic sample and all US non-federal, acute-care hospitals.

Results
As shown in Figure 1, in 2018, 65% of shared patients received care from hospital pairs connected to the same HIE network. Of these, 30% were connected via HIOs only, 18% were connected via vendor networks only, and 17% were connected by both types of networks. For 13% of shared patients, the two hospitals in the pair were connected to different networks (i.e., not able to electronically share patient data with each other). For 20% of shared patients, only one hospital in the pair participated in a network and for 2% neither hospital participated in a network.

Shared patients were more likely to be treated by hospitals on the same network when the pair shared a higher number of patients. Shared patients were less likely to be treated by hospitals on the same network when: 1) one member of the pair had a small number of other hospital partners while the other had a high number; 2) ownership
of one or both hospitals was for-profit, or both were government owned; and 3) both hospitals in the pair were located in moderately or highly competitive markets.

![Figure 1. Percent of Patients Shared by Two Hospitals, by Connectivity Status of the Two Hospitals (2018)](image)

**Discussion**

While national measures suggest that approximately 80% of hospitals participate in a HIE network\(^1\), only 65% of shared patients are treated by hospitals on the same HIE network. Targeted efforts to encourage lagging hospitals to join HIE networks could lead to exponential increases in connectivity for the 20% of shared patients in which only one hospital participates in any HIE network. More broadly, the federal Trusted Exchange Framework and Common Agreement, a voluntary national exchange framework that is still being finalized\(^2\), offers a logical policy vehicle to close the gaps we identify by explicitly considering how to prioritize participation based on shared patient volume. For example, the Recognized Coordinating Entity could identify hospitals and networks that, if connected, would most substantially increase connectivity based on shared patient volume. They could then share this list with the Qualified Health Information Networks (the networks that will be connected under TEFCA) to guide their recruitment and onboarding efforts. Our measures indicate that this explicit focus is necessary in order for ongoing public and private efforts aimed at connecting disparate HIE approaches to result in large gains in connectivity.

**References**

The Relationship between Multiple Methods used to Obtain Health Information and Use of Outside Information by Hospitals

Jordan Everson, PhD, MPP\textsuperscript{1} and Vaishali Patel, PhD\textsuperscript{1}

\textsuperscript{1}Office of the National Coordinator for Health IT at HHS

Introduction

Interoperable Health information exchange (HIE) technologies have developed such that healthcare providers can choose among several methods to obtain electronic patient health information from providers outside of their organization.\textsuperscript{1} However, in general each method does not facilitate exchange with providers using other methods or meet all use cases. In consequence, providers have adopted multiple methods to obtain and share information from outside providers, with the median hospital having adopted four methods in 2017.\textsuperscript{2}

Despite the prevalence of multiple methods of HIE, little is known about how these methods inter-relate and are associated with information availability and use by providers.\textsuperscript{3} Better characterizing how these methods are adopted, used, and relate to overall information availability could help guide health system strategy related to selecting strategies to engage in HIE and policy initiatives such as the 21\textsuperscript{st} Century Cures Act: The Trusted Exchange Framework and Cooperative Agreement, which is aimed at creating a 'single on-ramp' for information exchange by linking networks together such that participation in one network implies participation in all.\textsuperscript{4} We therefore sought to characterize how commonly each method of obtaining health information was used, how each method related to each other and how adoption of each method related to the likelihood that health information was routinely available to providers and often used.

Methods

We performed cross-sectional analysis of the 2019 American Hospital Association Information Technology Supplement to assess relationships between seven different electronic methods hospitals used to receive or find information from outside providers. Four methods involve different types of intermediaries or networks that facilitate exchange with many other hospital and providers. The other 3 methods enable one-to-one connections between organizations. In multivariable analysis controlling for hospital characteristics, we measured the relationship between adoption of each method and the likelihood that other methods were adopted and used, that providers had access to outside information and that providers used that information routinely. We report relative risk ratios rather than odds ratios for ease of interpretation.

Results

Single EHR vendor networks and national networks were the most commonly used method by those that had adopted each, with 63.8\% (95\% CI: 61.4\%-66.2\%) and 55.8\% (95\% CI: 53.4\%-58.2\%) of adopters reporting often using the method, respectively. Though most frequently adopted, provider portals were only reported as often used by 34.6\% (95\% CI: 32.5\%-36.7\%) of respondents.

In multivariate analysis, hospitals with regional HIEs were much more likely to also obtain information through Direct Messaging via a health information service provider (HISP) (RR=2.24; 95\% CI: 1.98-2.52; Figure 2); hospitals participating in national networks were modestly more likely to use Direct (RR=1.31; 95\% CI: 1.15-1.50) and hospitals with single vendor networks were only slightly more likely to use Direct (RR=1.13; 95\% CI: 1.05-1.21). Similarly, hospitals that obtained information through Direct Messaging via a HISP were much more likely to also obtain information through a regional HIE (RR=1.68; 95\% CI: 1.55-1.82), and hospitals in national networks were also more likely to use an HIE (RR=1.45; 95\% CI: 1.35-1.56). Hospitals with single vendor networks were again only slightly more likely to use an HIE (RR=1.07; 95\% CI: 1.00-1.14). Associations between the three methods without an intermediary and either Direct Messaging or HIE use were small.

Finally, hospitals that participated in Direct Messaging via a HISP, regional HIEs, and single vendor networks were more likely to participate in a national network (RR=1.48; 95\% CI: 1.34-1.64; RR=2.18; 95\% CI: 1.87-2.54; RR=1.64; 95\% CI: 1.50-1.80). Associations between the three methods without an intermediary and use of national networks were small.

All four methods of obtaining information with intermediaries were associated with greater likelihood of
Relative risk for Direct messaging via HISPs was 1.39 (95% CI: 1.25-1.54); for HIOs the relative risk was 1.19 (95% CI: 1.07-1.32); for single vendor networks the relative risk was 1.37 (95% CI: 1.25-1.49); and for national networks was 1.28 (95% CI: 1.17-1.40). These risk ratios are not statistically distinguishable (p>0.05 in all comparisons). However, only the single vendor networks and national networks were associated with greater risk of information being used by providers. The association between single vendor networks and use of information (Odds Ratio= 4.28; 95% CI=3.53-5.18) were substantial and larger than the association between national networks and use of information (Odds Ratio=1.87; 95% CI=1.59-2.20).

**Figure 2. Relationship between Methods of Obtaining Information**

Discussion
Specific methods of information exchange were highly related, indicating that these methods may be complementary or facilitate exchange by each other. As technology continues to evolve, and the Trusted Exchange Framework and Common Agreement facilitates exchange across networks, it will be important to assess whether these methods become more closely tied together or if the number of methods used decrease. Our findings also highlight important differences between available technologies. While all four HIE methods involving intermediaries were associated with greater availability of information, adoption of a regional HIE was not associated with more frequent use of that information. This reinforces concerns about the usability of data provided via these methods.5

**References**

Collecting and Sharing ECG Data with Privacy Protection

Liyue Fan PhD1, Zeyun Wu2, Luca Bonomi PhD3
1Department of Computer Science, University of North Carolina at Charlotte, NC
2UCSD Department of Electrical and Computer Engineering, La Jolla, CA
3UCSD Health Department of Biomedical Informatics, La Jolla, CA

Introduction. Advances in mobile sensor technology (e.g., wearable devices, smartphones) are revolutionizing the way in which electrocardiography (ECG) data are collected, enabling high-quality and fine-grained data to be gathered directly from the individual’s personal device1. Integrating these data into the clinical decision process has the potential for developing better interventions for cardiovascular disease and improving patient care2. However, there are significant privacy concerns in collecting and sharing ECG data. Data collected by sensing devices may contain patterns that could reveal sensitive information about individuals (e.g., heart rate and mobility patterns may be used to infer an individual’s activity). As a result, privacy methods are needed to protect data at the time of data collection. Furthermore, several studies have demonstrated that ECG data could be used as biometric to identify individuals3,4. As an example, Biel et al.4 have shown that carefully selected features from ECG could be used to achieve 100% re-identification rate on a dataset with 20 patients. As a result, simple de-identification methods that rely on removing personal health information (e.g., SSN), do not provide adequate privacy protection. To mitigate these privacy risks, current solutions for ECG data rely on encryption-based methods5. While these methods have shown promising results, their end-to-end encryption paradigm allows only a small number authorized health professionals to access the raw data. Additionally, encryption solutions may still be vulnerable in the presence of an informed adversary (e.g., in genomic data sharing applications6). Therefore, to facilitate the collection and sharing of ECG data, it is imperative to develop novel privacy technologies that address individual’s privacy concerns while enabling a broad access to the collected data.

In this on-going work, we examine how recent formal privacy models can benefit ECG data-driven applications. Specifically, we consider: metric privacy7 and local differential privacy8, which are extensions of the popular differential privacy model. These privacy models have shown to provide provable privacy protection for individual-level data and have been recently deployed by IT companies (e.g., Google, Apple) to provide privacy in IoT settings. Building on these rigorous privacy models, we develop privacy-protecting solutions for individual-level ECG data sanitization that can be deployed on each individual’s wearable device, protecting data as they are collected. We conduct empirical evaluations to understand the tradeoff between usability and privacy protection for clinical predictive tasks. Additionally, we also assess the privacy protection provided by these models when the collected individual data are first aggregated and then shared for secondary use (e.g., ECG data published for targeted cardiovascular studies). In such a setting, we consider an informed adversary who may use external information to infer the presence of a target individual in the study (i.e., membership inference).

Our preliminary evaluations on a real-world ECG data show that rigorous privacy models (e.g., metric privacy) for individual-level data can be adapted to provide privacy at the point of data collection, thus addressing individual’s privacy concerns and facilitating data participation. Additionally, these privacy models are successful in reducing the membership inference risk when the aggregate-level data are shared for secondary use.

Method. In this on-going work, we study the applicability of two recent privacy models (i.e., metric privacy and local differential privacy) in the ECG data collection and sharing setting. Specifically, we focus on: (1) privacy-utility trade-off in collecting individual-level ECG data for clinical predictive tasks, and (2) practical privacy risks when these data are aggregated and shared for secondary use.

Individual-level ECG data sanitization. In our work, we consider the ECG being collected as time-series data. To protect privacy during data collection, we design data sanitization methods that obfuscate the original ECG data as they leave the individual’s wearable devices, enabling individual to share a sanitized version of their ECG data with a central (not necessarily trusted) authority without revealing sensitive information. For example, our metric privacy method achieves privacy by first representing the time-series using the Discrete Cosine Transform (DCT), and then by carefully sampling the DCT domain a new set of coefficients are used to generate a sanitized time-series. Intuitively, under this privacy model, an individual can generate a sanitized time-series $\hat{s}$ that resembles the original ECG data $s$, for which the amount of information revealed by $\hat{s}$ about $s$ is controlled by the privacy parameter $\epsilon$. Each individual can increase the privacy protection by setting smaller values of $\epsilon$. Using this privacy paradigm, we study the impact of the user’s privacy preferences (e.g., values of the privacy parameter $\epsilon$) on the quality of the collected data for predictive analytics.

Privacy risk evaluation. To quantify the privacy protection provided by these privacy models when data are aggregated and shared for secondary use, we conduct empirical privacy risk evaluations to assess the membership inference risk. In this preliminary study, we consider an adversary who may have access to partial information of a target individual (e.g., few readings from the target). This assumption is motivated by the fact that individuals may inadvertently reveal part of the recorded data (e.g., by sharing the data on web portal provided by apps on their mobile devices) or their devices may be compromised. The goal of the adversary is to determine whether the target is in the sanitized ECG dataset, potentially learning sensitive information about the target (e.g., diagnosed with left ventricular dysfunction). We simulate such an attack, in which we control the number of readings from the original target ECG time-series data
known to the adversary and are used to re-identify the target in the shared data. For re-identification, we perform a similarity match between readings and the time-series using the Dynamic Time Warping distance (DTW). This similarity measure allows a temporal stretch in matching time-series. Furthermore, the attack is successful if the target’s time-series is identified in the top-5 nearest series to the target’s known readings (i.e., TPR@top5). Higher values of TPR indicate higher privacy risk.

**Results.** Our preliminary evaluations are conducted on the ECG200 dataset with 200 ECG time series\(^\text{10}\). Each series contains 96 samples and has been normalized to zero mean and unit variance. The dataset is labeled, where patients belong to two classes: normal heartbeat and Myocardial Infarction.

To evaluate the usability of the sanitized data, we considered a classification task, where a SVM classifier is used to classify patients as either normal or with myocardial infarction. The overall data is divided into 80% training and 20% testing. The training set is sanitized using metric privacy. We then train the SVM classifier on the sanitized data and measured the accuracy on the test set. The accuracy results are reported in Table 1. We observe that the usability increases as the value of the privacy parameter increases (i.e., weaker privacy), and the results obtained on sanitized data are comparable to those obtained on the original data (0.87 accuracy for non-private approach) when \(\varepsilon \geq 5.0\).

Figure 1 reports the empirical privacy risk in terms of true positive rate (TPR) for both the original ECG data and sanitized data with metric privacy. Our preliminary evaluations show that in the non-private data, an adversary who has access to roughly 40% of the readings achieves 0.6 TPR (i.e., more than 60% of the times a target is included in the top-5 nearest time-series). By sanitizing the ECG data using metric privacy with \(\varepsilon = 2.0\), the TPR reduces to around 0.32, thus providing significant privacy protection against this type of membership inference attack while preserving data usability.

**Discussion.** In this work, we presented our preliminary investigation on the applicability of recent formal privacy models to provide privacy protection for individual-level ECG data. Evaluations using the metric privacy model have shown that the obfuscated ECG data collected from each individual can be used to build predictive models for cardiovascular diseases, and the aggregated data can be broadly shared for secondary use. For the time of the conference, we will include our results using local differential privacy and will test our methods on larger datasets (e.g., MIMIC-III). Overall, our privacy study provides important insights on the development of privacy-protecting pipelines for collecting individual-level data and making them available for secondary use, which could facilitate emerging health applications (e.g., telemedicine and personalized medicine).

**References.**

Radiology Dictation Errors with COVID-19 Protective Equipment: Does Wearing a Surgical Mask Increase the Error Rate When Using Speech-Recognition?

Abiola Femi-Abodunde, MD 1, Kristen Olinger, MD 1, Lauren MB Burke, MD 1, Thad Benefield, MS 1, Ellie R. Lee, MD 1, Katrina McGinty, MD 1, Benjamin M. Mervak, MD 1,2

1 - Department of Radiology, University of North Carolina School of Medicine, Chapel Hill, NC
2 - Corresponding Author

Introduction: COVID-19 has resulted in numerous changes to the daily practice of medicine, including the use of face masks. One study demonstrated a negative impact of personal protective equipment (PPE) on interpersonal healthcare communication in a clinical setting. Radiologists widely use speech-recognition dictation software to generate radiology reports, although transcription is imperfect even in optimal situations. Masks have a yet-unknown effect on accuracy of speech recognition. The purpose of this study is to determine the effect of surgical masks on the number and type of dictation errors in unedited radiology reports.

Methods: IRB review was waived for this prospective matched-pairs study without patient data. A power analysis was conducted. Model radiology reports (n=40) simulated those typical for an academic medical center, divided evenly between Magnetic Resonance (MR), Computed Tomography (CT), Ultrasound (US), and X-ray (XR). Six radiologists used speech-recognition software to dictate from model reports with and without a surgical mask. Models and dictations were manually compared, and errors classified by type and severity. A statistical model compared error rates using a t-test. A sensitivity analysis was performed excluding a trainee with an exceptionally large number of errors.

Results: Error rates (per 1000 words) were 21.7 ± 4.9 when unmasked and 27.1 ± 2.2 when masked (25% increase, adjusted p<0.0001). A sensitivity analysis in a group with fewer errors revealed error rates of 16.9 ± 1.9 when unmasked and 20.1 ± 2.2 when masked (19% increase, adjusted p=0.054). Single incorrect word errors were the most common (64-66%), and most errors were minor (56-58%). MR and XR reports had higher error rates than CT reports. For some participants, MR reports had significantly more errors than US reports.

Discussion: Wearing a mask was found to result in more overall errors and particularly affected some radiologists. However, most errors were minor single incorrect words and, while a nuisance, are unlikely to affect clinical care.

References:

### Error Rates - All Participants

<table>
<thead>
<tr>
<th>Error Type (Severity)</th>
<th>Without Mask *</th>
<th>With Mask *</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Incorrect Word (All)</td>
<td>14.3 ± 2.7</td>
<td>15.9 ± 2.9</td>
<td>0.11</td>
</tr>
<tr>
<td>Incorrect Word (Minor)</td>
<td>7.2 ± 1.2</td>
<td>8.5 ± 1.4</td>
<td>0.067</td>
</tr>
<tr>
<td>Incorrect Word (Moderate)</td>
<td>2.6 ± 0.9</td>
<td>2.2 ± 0.8</td>
<td>0.27</td>
</tr>
<tr>
<td>Incorrect Word (Major)</td>
<td>3.9 ± 1.0</td>
<td>5.0 ± 1.3</td>
<td>0.044</td>
</tr>
<tr>
<td>Missing Word (All)</td>
<td>3.5 ± 0.9</td>
<td>4.3 ± 1.1</td>
<td>0.049</td>
</tr>
<tr>
<td>Missing Word (Minor)</td>
<td>2.7 ± 0.6</td>
<td>3.0 ± 0.7</td>
<td>0.37</td>
</tr>
<tr>
<td>Missing Word (Moderate)</td>
<td>0.3 ± 0.2</td>
<td>0.6 ± 0.4</td>
<td>0.001</td>
</tr>
<tr>
<td>Missing Word (Major)</td>
<td>0.4 ± 0.1</td>
<td>0.4 ± 0.1</td>
<td>0.99</td>
</tr>
<tr>
<td>Additional Word (All)</td>
<td>6.7 ± 0.9</td>
<td>2.0 ± 0.4</td>
<td>0.39</td>
</tr>
<tr>
<td>Additional Word (Minor)</td>
<td>1.1 ± 0.2</td>
<td>0.1 ± 0.2</td>
<td>0.58</td>
</tr>
<tr>
<td>Additional Word (Moderate)</td>
<td>0.3 ± 0.08</td>
<td>0.5 ± 0.1</td>
<td>0.27</td>
</tr>
<tr>
<td>Additional Word (Major)</td>
<td>0.02 ± 0.02</td>
<td>0.07 ± 0.03</td>
<td>0.27</td>
</tr>
<tr>
<td>Erroroneous Phrase (Major)</td>
<td>0.6 ± 0.3</td>
<td>0.8 ± 0.3</td>
<td>0.29</td>
</tr>
<tr>
<td>Error in Term of Negation (Major)</td>
<td>0.1 ± 0.05</td>
<td>0.2 ± 0.1</td>
<td>0.018</td>
</tr>
<tr>
<td>Numeric Errors (All)</td>
<td>11.1 ± 2.6</td>
<td>15.2 ± 3.2</td>
<td>0.0002</td>
</tr>
<tr>
<td>All Minor Errors</td>
<td>4.1 ± 1.3</td>
<td>4.3 ± 1.4</td>
<td>0.63</td>
</tr>
<tr>
<td>All Moderate Errors</td>
<td>5.6 ± 1.6</td>
<td>7.3 ± 2.0</td>
<td>0.008</td>
</tr>
<tr>
<td>All Major Errors</td>
<td>21.7 ± 4.9</td>
<td>27.1 ± 6.0</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>All Errors</td>
<td>21.7 ± 4.9</td>
<td>27.1 ± 6.0</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

### Error Rates - Attending Subgroup

<table>
<thead>
<tr>
<th>Error Type (Severity)</th>
<th>Without Mask *</th>
<th>With Mask *</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Incorrect Word (All)</td>
<td>10.9 ± 2.7</td>
<td>13.3 ± 1.4</td>
<td>0.054</td>
</tr>
<tr>
<td>Incorrect Word (Minor)</td>
<td>5.9 ± 0.9</td>
<td>7.5 ± 1.0</td>
<td>0.066</td>
</tr>
<tr>
<td>Incorrect Word (Moderate)</td>
<td>1.7 ± 0.4</td>
<td>1.6 ± 0.4</td>
<td>0.87</td>
</tr>
<tr>
<td>Incorrect Word (Major)</td>
<td>3.2 ± 0.9</td>
<td>4.1 ± 1.1</td>
<td>0.21</td>
</tr>
<tr>
<td>Missing Word (All)</td>
<td>2.9 ± 0.6</td>
<td>3.0 ± 0.7</td>
<td>0.86</td>
</tr>
<tr>
<td>Missing Word (Minor)</td>
<td>2.3 ± 0.5</td>
<td>2.2 ± 0.5</td>
<td>0.96</td>
</tr>
<tr>
<td>Missing Word (Moderate)</td>
<td>0.2 ± 0.2</td>
<td>0.3 ± 0.2</td>
<td>0.28</td>
</tr>
<tr>
<td>Missing Word (Major)</td>
<td>0.3 ± 0.1</td>
<td>0.3 ± 0.1</td>
<td>0.99</td>
</tr>
<tr>
<td>Additional Word (All)</td>
<td>1.4 ± 0.3</td>
<td>1.7 ± 0.3</td>
<td>0.74</td>
</tr>
<tr>
<td>Additional Word (Minor)</td>
<td>1.1 ± 0.3</td>
<td>1.1 ± 0.3</td>
<td>&lt;0.99</td>
</tr>
<tr>
<td>Additional Word (Moderate)</td>
<td>0.3 ± 0.08</td>
<td>0.4 ± 0.1</td>
<td>0.61</td>
</tr>
<tr>
<td>Additional Word (Major)</td>
<td>0.02 ± 0.02</td>
<td>0.06 ± 0.04</td>
<td>0.61</td>
</tr>
<tr>
<td>Erroroneous Phrase (Major)</td>
<td>0.4 ± 0.2</td>
<td>0.5 ± 0.2</td>
<td>0.49</td>
</tr>
<tr>
<td>Error in Term of Negation (Major)</td>
<td>0.08 ± 0.04</td>
<td>0.1 ± 0.05</td>
<td>0.74</td>
</tr>
<tr>
<td>Numeric Errors (All)</td>
<td>0.7 ± 0.2</td>
<td>0.9 ± 0.2</td>
<td>0.74</td>
</tr>
<tr>
<td>Incorrect Measurement (Moderate)</td>
<td>0.1 ± 0.08</td>
<td>0.2 ± 0.1</td>
<td>0.74</td>
</tr>
<tr>
<td>Incorrect Image Number (Moderate)</td>
<td>0.2 ± 0.1</td>
<td>0.2 ± 0.1</td>
<td>&lt;0.99</td>
</tr>
<tr>
<td>Incorrect Date/Time (Moderate)</td>
<td>0.2 ± 0.1</td>
<td>0.3 ± 0.1</td>
<td>0.74</td>
</tr>
<tr>
<td>Punctuation Errors (Minor)</td>
<td>0.3 ± 0.2</td>
<td>0.3 ± 0.2</td>
<td>0.74</td>
</tr>
<tr>
<td>All Minor Errors</td>
<td>9.8 ± 1.1</td>
<td>11.3 ± 1.3</td>
<td>0.2</td>
</tr>
<tr>
<td>All Moderate Errors</td>
<td>2.8 ± 0.6</td>
<td>3.1 ± 0.6</td>
<td>0.74</td>
</tr>
<tr>
<td>All Major Errors</td>
<td>4.3 ± 1.1</td>
<td>5.6 ± 1.3</td>
<td>0.11</td>
</tr>
<tr>
<td>All Errors</td>
<td>16.9 ± 1.9</td>
<td>20.1 ± 2.2</td>
<td>0.054</td>
</tr>
</tbody>
</table>

### Error Rates - Modality

<table>
<thead>
<tr>
<th>Modality</th>
<th>Error Rate * (All Participants)</th>
<th>Error Rate * (Attending Subgroup)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CT</td>
<td>19.7 ± 4.5</td>
<td>15.3 ± 2.0</td>
</tr>
<tr>
<td>MRI</td>
<td>29.3 ± 6.5</td>
<td>21.1 ± 2.6</td>
</tr>
<tr>
<td>US</td>
<td>21.9 ± 5.0</td>
<td>16.6 ± 2.4</td>
</tr>
<tr>
<td>XR</td>
<td>26.4 ± 6.1</td>
<td>22.2 ± 3.3</td>
</tr>
</tbody>
</table>

### Comparison

<table>
<thead>
<tr>
<th>Comparison</th>
<th>P-Value (All Participants)</th>
<th>P-Value (Attending Subgroup)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CT vs MRI</td>
<td>&lt;0.0001</td>
<td>0.043</td>
</tr>
<tr>
<td>CT vs US</td>
<td>0.28</td>
<td>0.89</td>
</tr>
<tr>
<td>CT vs XR</td>
<td>0.009</td>
<td>0.043</td>
</tr>
<tr>
<td>MRI vs US</td>
<td>0.003</td>
<td>0.14</td>
</tr>
<tr>
<td>MRI vs XR</td>
<td>0.28</td>
<td>0.71</td>
</tr>
<tr>
<td>US vs XR</td>
<td>0.15</td>
<td>0.14</td>
</tr>
</tbody>
</table>

1365
Identifying Communication Behavior Indicators in Secure Messages

Dezon K Finch, PhD¹, Lina Bouayad, PhD², Timothy P Hogan, PhD³, Sarah L Cutrona, MD, MPH³,⁴, Byron Wallace, PhD⁵, Stephen L Luther, PhD¹,⁶, Bridget M Smith, PhD⁷, Stephanie L Shimada, PhD⁴,⁷,⁸

¹James A Haley Veterans Hospital, Tampa, FL; ²Florida International University, Miami, FL; ³VA Bedford Healthcare System, Bedford, MA; ⁴Department of Population and Quantitative Health Sciences, University of Massachusetts Medical School, Worcester, MA; ⁵Northeastern University, Boston, MA; ⁶University of South Florida, College of Public Health, Tampa FL; ⁷Edward Hines Jr VA Hospital, Chicago, IL; ⁸Department of HLPM, Boston University School of Public Health, Boston, MA

Introduction

Communication between patients and healthcare providers can impact patient-provider relationships, yet capturing and analyzing these interactions is time-consuming and costly, relying on video- or audio-recordings and laboriously coding them. We sought to determine whether a machine learning algorithm could be developed to more efficiently identify important communication behaviors within asynchronous secure messages (SM) written by healthcare team members. We adapted the Roter Interaction Analysis System¹ (RIAS) method to identify Communication Behavior Indicators (CBIs) in provider messages within SM threads exchanged between patients and clinical team members.

Methods

Secure messaging is a part of the Veteran’s Health Administration’s (VA’s) My HealtheVet patient portal. Secure messages stored as progress notes within the Electronic Health Record are available as Text Integration Units (TIUs) in the VA Corporate Data Warehouse (CDW). To identify CBIs in SM threads, we first created a human annotated reference set of threads to develop and evaluate a Natural Language Processing (NLP) system. To ensure the richest possible messages, we chose the thread with the highest number of messages for each patient. Based on the facility usage of SM, patient’s threads were proportionally sampled and stratified based on gender and number of patients. Since the Veteran population is predominantly male, we oversampled females to ensure adequate representation. We adapted a limited set of RIAS codes appropriate for SMs and created a coding manual for use by the team. Each message thread was annotated by two clinicians and adjudicated by a third expert. Agreement was evaluated for each of the in 36 sets of 250 threads so agreement could be checked as annotations progressed. Agreement varied but remained above .7 on average for the entire corpus. Targeted CBIs are show below in Table 1.

Table 1. Targeted Team CBIs

<table>
<thead>
<tr>
<th>CBI Name</th>
<th>Definition</th>
<th>Prevalence</th>
</tr>
</thead>
<tbody>
<tr>
<td>InfoGiving</td>
<td>Statements/Opinions regarding the patients’ health</td>
<td>52.53%</td>
</tr>
<tr>
<td>GenericConnect</td>
<td>Greetings, friendly statements or gestures, formal greetings, goodbyes.</td>
<td>16.43%</td>
</tr>
<tr>
<td>Instruction</td>
<td>Statements explaining what they are to do and/or how they are to do it.</td>
<td>12.84%</td>
</tr>
<tr>
<td>InfoSeeking</td>
<td>Any questions pertaining to the patient’s health</td>
<td>6.56%</td>
</tr>
<tr>
<td>ExpressConcern</td>
<td>Statements that something is serious, worrisome or distressing</td>
<td>2.65%</td>
</tr>
<tr>
<td>PartnershipBuild</td>
<td>Statements that build alliance with help, support and decision-making</td>
<td>2.45%</td>
</tr>
<tr>
<td>EncouragesConverse</td>
<td>Encourages the patient to continue the interaction with the team.</td>
<td>1.92%</td>
</tr>
<tr>
<td>GiveReassurance</td>
<td>Statements indicating optimism or encouragement.</td>
<td>1.74%</td>
</tr>
<tr>
<td>Counsels</td>
<td>Statements to persuade, influence or change the patient’s behavior</td>
<td>1.19%</td>
</tr>
<tr>
<td>AskForOpinion</td>
<td>Questions that ask for the patients’ opinion relating to their healthcare.</td>
<td>0.69%</td>
</tr>
<tr>
<td>SpecificConnect</td>
<td>Remarks specific to the patient, to establish a more personal relationship.</td>
<td>0.59%</td>
</tr>
<tr>
<td>ExpressEmpathy</td>
<td>Statements that recognize the emotional state of the patient.</td>
<td>0.13%</td>
</tr>
<tr>
<td>CheckUnderstand</td>
<td>Checks with the patient to see if information has been understood</td>
<td>0.06%</td>
</tr>
<tr>
<td>LegitimizeEmotion</td>
<td>Statements that the patient’s emotions, actions, or thoughts are normal</td>
<td>0.06%</td>
</tr>
</tbody>
</table>
After annotation was complete, we performed data cleaning and transformation on the threads to produce a dataset ready for analysis. The author of each message in the thread was assigned based on the “to” and “from” fields, and the datasets was created for analysis. Each sentence, phrase, or line of the body of the message was assigned a value encoding the corresponding annotations. Every line was assigned a value representing a targeted CBI. To improve incidence CBIs, we combined them into groups of related CBIs based on RIAS CBI categories still meaningful to the study. The resulting grouped CBIs and the component CBIs are listed below in Table 2. We removed lines that were not written in English, as well as blank lines and those containing only numbers.

Table 2. New Targeted CBI Groups

<table>
<thead>
<tr>
<th>Group CBIs</th>
<th>Previous Team CBIs Combined</th>
<th>Prevalence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Counsel</td>
<td>Counsel, InfoGive, InfoSeek, Instruction</td>
<td>63.02%</td>
</tr>
<tr>
<td>Connect</td>
<td>GennericConnect, SpecialConnect</td>
<td>28.04%</td>
</tr>
<tr>
<td>Build</td>
<td>PartnershipBuild, EncouragesConverse, CheckUnderstanding, AskForOpinion</td>
<td>4.79%</td>
</tr>
<tr>
<td>Empathy</td>
<td>Empathy, Legitimize, Reassure, ExpConcern</td>
<td>4.15%</td>
</tr>
</tbody>
</table>

The 9000 threads resulted in a total of 138,294 lines of text from the bodies of the team messages. We held out the lines of 2000 of the threads as a validation set, leaving 7000 for development. During model development the remaining data were split for into 80/20 training and test sets. We considered two multi-class models: logistic regression using the sklearn package and a feed-forward neural network, namely a multilayer perceptron (MLP). We preprocessed the text by tokenizing and removing punctuation; we then padded instances such that they uniformly comprise 100 tokens. Models were developed in Python using the Keras package with Tensorflow as a backend.

We performed hyperparameter search over the number of epochs, size of vocabulary, activation types, and other hyperparameters. We considered several methods to counteract “class imbalance” in the data, including class weighting and balanced batches. We also considered using multiple MLP models in tandem, combining the prediction probabilities of eleven separate models and deriving a new decision point to determine class membership.

Results

The results of the final best models are presented below. The logistic regression fared as well as the MLP. The overall results were better than our initial models on the individual CBIs. Ten of the original CBIs were represented in the data at under 3% of the lines of text and produced f-scores of under .5. For the combined CBIs performance on Counsel and Connection were acceptable, performance on Empathy and Build were weak. Precision for the weak performers were both over 0.7 which is desirable since for this study, since we are more concerned that the instances identified as positive be a true positive.

Table 3. Final Model Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Multilayer Perceptron</th>
<th>Logistic Regression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group CBIs</td>
<td>precision recall f1-score</td>
<td>precision recall f1-score</td>
</tr>
<tr>
<td>Build</td>
<td>0.76 0.20 0.31 0.75</td>
<td>0.51 0.60</td>
</tr>
<tr>
<td>Counsel</td>
<td>0.88 0.96 0.92 0.92</td>
<td>0.97 0.94</td>
</tr>
<tr>
<td>Connection</td>
<td>0.88 0.92 0.90 0.95</td>
<td>0.93 0.94</td>
</tr>
<tr>
<td>Empathy</td>
<td>0.71 0.02 0.03 0.78</td>
<td>0.55 0.65</td>
</tr>
</tbody>
</table>

Conclusion

To our knowledge, this study represents the first attempt to automatically extract CBIs from secure messages. The targeted CBIs were abstract, and some of the CBIs were rare (occurring in ~3% of the lines of text). Combining CBIs into related groups as new CBIs provided better targets in the PACT’s message data and we were able to achieve high F-scores for two of the four CBIs and acceptable precision on the other two. Future work includes examining associations between CBIs and patient reported satisfaction scores at the patient and facility level.

References

High-throughput sequencing of SARS-CoV-2 in wastewater provides insights into circulating variants

Rafaela S Fontenele¹, Simona Kraberger, PhD¹, James Hadfield, PhD², Erin M Driver, PhD³, Devin Bowes¹, LaRinda A Holland¹, Temitope O C Faley, PhD¹, Sangeet Adhikari¹, Rahul Kumar, PhD¹, Rosa Inchausti³, Wydale K Holmes³, Stephanie Deitrick³, Philip Brown³, Darrell Duty³, PhD³, Aruni Bhatnagar¹, Ray A Yeager, PhD, MPH¹, Rochelle H Holm⁴, Natalia Hoogesteijn von Reitzenstein, PhD³, Elliott Wheeler⁵, Kevin Dixon⁵, Tim Constantine⁵, Melissa A Wilson, PhD¹, Efrem S Lim, PhD¹, Xiaofang Jiang, PhD⁴, Rolf U Halden, PhD¹,⁷, Matthew Scotch, PhD, MPH¹, Arvind Varsani, PhD¹

¹Arizona State University, Tempe, AZ; ²Fred Hutch Cancer Research Center, Seattle, WA; ³City of Tempe, AZ; ⁴University of Louisville, KY; ⁵Jacobs Engineering Group Inc., Dallas, TX; ⁶National Library of Medicine, Bethesda, MD; ⁷OneWaterOneHealth, Tempe, AZ

Introduction

Wastewater-based epidemiology (WBE) and genomic epidemiology represent non traditional forms of public health surveillance that have been leveraged during the SARS-CoV-2 pandemic. The former is possible since SARS-CoV-2 is shed through bodily excretions while the latter uses high-throughput sequencing (HTS) of SARS-CoV-2 genomes to highlight transmission events and the diversity of circulating variants. Since both symptomatic and asymptomatic individuals contribute to wastewater, we hypothesized that pooled samples of population-wide excreta can provide a more comprehensive picture of SARS-CoV-2 genomic diversity in a community than clinical sequencing alone.

Methods

We collected 24-hr composite samples of untreated wastewater using automated samplers from locations in 11 states from 7th April to 16th June 2020. We filtered aliquots of 150 ml of each composite sample via a 0.45 µm polyethersulfone (PES) filter and then subsequently through a 0.2 µm PES filter. For each sample, we used a 200 µl aliquot to extract total RNA. To determine the presence of SARS-CoV-2 in wastewater samples, we used the extracted RNA in a reverse transcription-quantitative PCR (RT-qPCR) assay targeting the E gene, as designed and validated by Corman et al.¹. We generated cDNA from the total RNA from each sample and used 10 µl for Illumina sequencing libraries with the Swift Nomalase® Amplicon SARS CoV-2 Panel. The samples were subsequently normalized, pooled and sequenced at Psomagen on an Illumina HiSeq 2500 sequencer (2×100 paired-end option on 1 lane in rapid mode).

Figure 1: Novel SARS-CoV-2 SNVs (not yet detected in clinical-derived samples as of 17th June 2020) in the 52 wastewater samples. y-axis: number of samples with the SNV, x-axis: position of SNV in the SARS-CoV-2 genome.
Results

We collected 60 of our 91 samples (66%) in Arizona (9 locations located in Maricopa County, AZ), 12 (13%) were from 9 locations in Louisville, KY, and 19 (21%) from other states. The tiling PCR amplification enrichment process for the SARS-CoV-2 genome generated 341 amplicons covering ~99% of the genome. We detected 7,973 single nucleotide variants (SNVs) in the 52 samples. For these, we identified 548 novel SNVs, of which 469 were non-synonymous (not including nonsense mutations) and 79 were synonymous substitutions (Figure 1). Since we evaluated all variants regardless of frequency, some locations (as expected) had more than one possible variant (Figure 1). These 548 SNVs are distributed along the SARS-CoV-2 genome with three located in non-coding regions.

In Figure 2, we show our principal coordinate analysis (PCoA) results using nucleotide frequencies to evaluate the viral population diversity within and between samples. Overall, sequences from the 10 states were highly diverse, and those from the same state clustered closer together. An exception to this are the AZ sequences which are broadly distributed. Conversely, the samples from Louisville, KY are more tightly clustered despite sampling from several city locations over a two-month period. One factor to the differences in viral diversity could be that Tempe (where most of the AZ samples originated) includes one of the largest universities in the USA and is near a busy international airport.

Conclusion

This study highlights the use of WBE and HTS for analyzing the genomics of ongoing outbreaks of infectious diseases, such as SARS-CoV-2. This approach can provide novel information on SNVs which, when coupled with data derived from clinical cases, can help identify new emerging variants of importance within a population. These approaches can be used within a public health setting as an early warning tool to inform infectious disease mitigation measures.

References

Can ICD-11 Replace ICD-10-CM for Morbidity Coding in the U.S.?

Kin Wah Fung¹, MD, MS, MA; Julia Xu¹, MD, PhD; Shannon McConnell-Lamptey², MS, RHIA; Donna Pickett², MPH, RHIA; Olivier Bodenreider¹, MD, PhD

¹National Library of Medicine, National Institutes of Health, USA; ²National Center for Health Statistics, Centers for Disease Control and Prevention, USA

Introduction

The International Classification of Diseases (ICD) has been in use for collection of global health trends and statistics for over a century. ¹ Its latest version, ICD-11, was adopted in May 2019 and will be implemented in member countries of the World Health Organization (WHO) from January 2022. ² Due to specific requirements in some countries, over two dozen national extensions of ICD have been developed for past versions of ICD. In the U.S., the first version of the national extension known as Clinical Modification (CM) was ICD-9-CM released in 1979. According to the official documentation of ICD-9-CM, “the term "clinical" is used to emphasize the modification's intent: to serve as a useful tool in the area of classification of morbidity data for indexing of medical records, medical care review, and ambulatory and other medical care programs, as well as for basic health statistics. To describe the clinical picture of the patient, the codes must be more precise than those needed only for statistical groupings and trend analysis.” ³ The same practice of modifying the international ICD core for clinical purpose continued in ICD-10-CM, which replaced ICD-9-CM in 2015.

The main advantage of developing a U.S. national extension is the ability to add necessary detail under the framework of the international core to serve clinical and administrative (e.g., reimbursement) needs. Another advantage is that updates to the national extension can happen more frequently, as ICD-10-CM is updated yearly compared to the three-year cycle for ICD-10. However, there are potential drawbacks. Firstly, significant effort is involved in maintaining an extension. Secondly, there is usually a delay between the release of the international version and the national extension. Moreover, there can be incongruence between the national extension and the international core. In principle, everything in the Clinical Modification should be totally compatible with the parent system. However, some significant differences can be observed between ICD-10-CM and ICD-10. For example, the ICD-10 category E14 Unspecified diabetes mellitus is not present in ICD-10-CM, because diabetes mellitus of unspecified type is coded under E11 Type 2 diabetes mellitus by default. Another example is the addition to ICD-10-CM of a new category K68 Disorders of retroperitoneum that is not present in ICD-10.

ICD-11 has some new features not available in previous versions. Apart from the introduction of the foundation component, the most noticeable novel feature in ICD-11 is postcoordination. ⁴ Postcoordination is the combination of codes to represent new meaning - a powerful and efficient way to expand the coverage, expressivity and granularity of a terminology. Towards this end, ICD-11 offers 14,500 extension codes for postcoordination. This new capability, together with the considerable increase in the number of codes - 4,015 (37.9%) more codes than ICD-10, may lead one to question whether it is still necessary to develop a Clinical Modification for ICD-11. The recommendations from the National Committee on Vital and Health Statistics (NCVHS) to the Secretary of the Department of Health and Human Services include research to determine whether ICD-11 can fully support morbidity classification in the U.S. without development of a U.S. clinical modification. ⁵

In a previous study, we examined a limited sample of ICD-10-CM codes to see how well they could be represented in ICD-11. ⁴ The objective of the present investigation is to do a comprehensive assessment of the feasibility of replacing ICD-10-CM with ICD-11 for morbidity coding, without creating a CM extension.

Methods

We used two sources of data to identify the most commonly used ICD-10-CM codes. First source was a full year of Medicare claims data in 2017. Since the Medicare population did not cover obstetric and pediatric codes, we supplemented it with data from three hospitals in Nebraska. In both sources, we identified the most commonly used ICD-10-CM codes that accounted for at least 60% of usage in each chapter. For each ICD-10-CM code, we identified the best matching ICD-11 code by using WHO’s online ICD-11 browser. ⁶ If the ICD-10-CM code’s meaning was not fully represented by the ICD-11 code, we would attempt postcoordination, as guided by the browser, to achieve full representation. Each ICD-11 recoding was done independently by two terminologists who are very knowledgeable in ICD-10-CM and ICD-11. All discrepancies were recorded and discussed until consensus was reached. Failure analysis was then performed for all codes that only achieved partial representation to determine the reason for failure and the type of missing information. Further review was carried out on the accompanying coding guidance - inclusion terms, exclusion terms and index entries - for potential conflicts between
the ICD-10-CM and ICD-11 codes e.g., the ICD-10-CM code had an inclusion which was an exclusion of the matching ICD-11 code.

Results

Based on both data sources, we identified altogether 962 unique ICD-10-CM codes required to cover 60% of patients, of which 943 codes were still active in 2021. Overall, of the 943 codes, 221 (23.4%) could be fully represented without postcoordination and 81 codes (8.6%) could be fully represented with postcoordination (Table). Concerning coding variability, agreement on the ICD-11 main codes was observed in 716 (75.9%) cases before discussion. Among these 716 cases, postcoordination was used by both terminologists in 253 cases, and they used the same postcoordination codes in 199 cases (78.7% agreement). Failure analysis showed that many of the partially represented codes were missing information related to episode of care, laterality, mode of drug exposure and trimester of pregnancy. With the addition of only nine extension codes in ICD-11 (three episodes of care, three trimesters of pregnancy and three modes of exposure), the proportion of codes that could be fully-postcoordinated will increase to 35.2%. Coding guidance review showed potential conflicts in 10% of the codes. For example, the ICD-10-CM code A41.9 Sepsis, unspecified organism was recoded as the ICD-11 code J140 Sepsis without septic shock. “Septicemia” was an inclusion for A41.9 but an exclusion for J140. In ICD-11, “septicemia” pointed to MA15 Microbiological findings in blood, blood-forming organs, or the immune system. Therefore, the ICD-11 code A41.9 was correct in the broader context of sepsis. However, in the special case of septicemia one should use MA15.

Table. Recoding commonly used ICD-10-CM codes in ICD-11

<table>
<thead>
<tr>
<th></th>
<th>ICD-11</th>
<th></th>
<th>ICD-11 (with minor enhancements)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Full representation without postcoordination</td>
<td>221</td>
<td>23.4%</td>
<td>221</td>
<td>23.4%</td>
</tr>
<tr>
<td>Full representation with postcoordination</td>
<td>81</td>
<td>8.6%</td>
<td>332</td>
<td>35.2%</td>
</tr>
<tr>
<td>Partial representation only</td>
<td>641</td>
<td>68.0%</td>
<td>390</td>
<td>41.4%</td>
</tr>
<tr>
<td>Total</td>
<td>943</td>
<td>100.0%</td>
<td>943</td>
<td>100.0%</td>
</tr>
</tbody>
</table>

Discussion

Without postcoordination, only 23.4% of ICD-10-CM codes are fully represented in ICD-11. However, this can be increased to 58.6% with postcoordination and some minor enhancements. Considering that only 24.3% of ICD-9-CM codes have exact matches to ICD-10-CM codes (based on the 2016 General Equivalence Maps), migrating from ICD-10-CM to ICD-11 appears to be less disruptive than the ICD-9-CM to ICD-10-CM transition. However, postcoordination increases coding complexity and will impact tooling, coder education and coding variability. Replacing ICD-10-CM directly with ICD-11 will avoid the cost of maintaining a national extension and the potential divergence of meaning from the international core. Before embarking on the development of ICD-11-CM, serious consideration should be given to using ICD-11 for morbidity coding.

Acknowledgements

This research was supported by the Intramural Research Program of the NIH, National Library of Medicine.

References

Mining and characterizing long-COVID symptoms from self-reports on Reddit

Yao Ge, MS\textsuperscript{1}, Yuting Guo, MS\textsuperscript{1}, Abeed Sarker, PhD\textsuperscript{1}
\textsuperscript{1}Department of Biomedical Informatics, School of Medicine, Emory University, Atlanta, GA, United States

Introduction

Many patients continue to experience symptoms long after the acute phase of infection with the SARS-CoV-2 (COVID-19) virus,\textsuperscript{1} and currently, the persistence of symptoms 28 days after the diagnosis of COVID-19 infection is referred to as ‘post-acute COVID syndrome’, ‘long-haul COVID-19’ or ‘long-COVID’. A growing body of research is exploring long-COVID and the constellation of symptoms that patients experience following the acute phase of COVID-19 infection. However, there are still many unknowns associated with long-COVID, including the full spectrum of symptoms that patients experience, their progression, and their long-term manifestation. One potential source of information regarding long-COVID syndrome is social media, as many people, including healthcare professionals who had been infected with COVID-19, are reporting their experiences through this medium.\textsuperscript{2} In this study, we attempted to mine publicly available data from Reddit to discover and analyze the spectrum of symptoms self-reported by sufferers of long-COVID. Our objectives for this study were to extend a COVID-19 symptom lexicon built using Twitter data, deploy the extended lexicon to identify self-reported long-COVID symptoms from a specific forum on Reddit, analyze the distribution of symptoms and compare them with symptoms reported in recent studies.

Methods

We collected data for study from the subreddit /r/covidlonghaulers, which has emerged as the go-to forum for the discussion of long-COVID-related topics, and it is an information-rich source for obtaining self-reported long-COVID symptoms. As of 6\textsuperscript{th} June, 2021, the subreddit had over 14,000 subscribers and over 1200 threads, most of which focus on long-COVID experiences self-reported by users, their symptoms, and often the timelines of their symptoms. We first manually annotated a symptom lexicon from Reddit, grouping similar symptoms and mapping each symptom expression to a standard identifier in the unified medical language system (UMLS) using the National Center for Biomedical Ontology BioPortal.

To detect symptoms from free text, we applied an inexact matching method. Exact matching on social media free texts typically results in low recall due to the presence of nonstandard expressions and misspellings. To overcome this issue, in line with our past work, we searched through term sequence windows in the texts and computed the similarity of each text window with all entries in the meta-lexicon. Text windows that obtained similarities above a specific threshold with any entry in the lexicon were extracted and considered to be candidates for long-COVID symptoms. For an expression of $n$ terms, the term sequence windows were of the range $[n-1:n+2]$. A negation detection algorithm is then applied to search for the occurrence of negations in the same post as the candidate symptom and estimating if the symptom appears within the scope of the negation. The negation detection algorithm labels any term appearing within a 3 token window or before an end-of-sentence marker following the negation term (eg., a period) as negated. We grouped the symptoms per user and generated their frequency distributions for comparison.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure1.png}
\caption{Number of users reporting each of the top 20 symptoms and their percentages.}
\end{figure}
We evaluated the performance of the algorithm using the precision, recall and F1-score metrics. To estimate precision and recall, we selected a random sample of posts, including posts from which at least one symptom was automatically detected and posts from which our algorithm did not detect any.

Results

We collected 42,995 posts from 1220 threads and 4249 users. We reviewed posts associated with a total of 450 symptom expressions (i.e., the same post could be reviewed multiple times in association with different detected symptoms) along with 25 posts without any detected symptoms to estimate the performance of our approach (recall = 0.93, precision = 0.95, F1-score = 0.94). 1744 users expressed at least 1 non-negated symptom (2576 users without accounting for negations).

<table>
<thead>
<tr>
<th></th>
<th>This study (n=1744)</th>
<th>Huang et al.3 (n=1733)</th>
<th>Sykes et al.4 (n=134)</th>
<th>Orrù et al.5 (N=152)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fatigue</td>
<td>51.2</td>
<td>63</td>
<td>39.6</td>
<td>74.3</td>
</tr>
<tr>
<td>Ache &amp; pain</td>
<td>48.4</td>
<td>67†</td>
<td>51.5†</td>
<td>61.2†</td>
</tr>
<tr>
<td>Brain fog/confusion or dizziness</td>
<td>32.8</td>
<td>67†</td>
<td>25.4†</td>
<td>48.7</td>
</tr>
<tr>
<td>Dyspnea</td>
<td>28.9</td>
<td>N/A</td>
<td>59.7†</td>
<td>40.1</td>
</tr>
<tr>
<td>Fever/pyrexia</td>
<td>24.5</td>
<td>&lt;1</td>
<td>10.4</td>
<td>19.1</td>
</tr>
<tr>
<td>Headache</td>
<td>22.4</td>
<td>2</td>
<td>N/A</td>
<td>46.7</td>
</tr>
<tr>
<td>Insomnia/sleep disturbance</td>
<td>9.7</td>
<td>26</td>
<td>35.1</td>
<td>N/A</td>
</tr>
<tr>
<td>Chest pain</td>
<td>18.3</td>
<td>5</td>
<td>17.9</td>
<td>26.3</td>
</tr>
<tr>
<td>Cough</td>
<td>15.9</td>
<td>N/A</td>
<td>35.1</td>
<td>21.1</td>
</tr>
<tr>
<td>Palpitations</td>
<td>17.3</td>
<td>9</td>
<td>N/A</td>
<td>38.8</td>
</tr>
</tbody>
</table>

Table 1. Comparison of long-COVID symptom distributions (in percentage) identified in this study with those reported in recent literature for frequently-reported symptoms. Estimated values are marked with †.

Figure 1 presents the raw counts for the top 20 symptoms reported by the users and their relative frequencies as a percentage of the total number of users who reported at least 1 non-negated symptom. The most commonly reported long-COVID symptoms are anxiety/stress and related mental health symptoms. The next 4 most commonly reported symptoms are fatigue, general & body ache & pain, confusion/disorientation and dyspnea. Pyrexia, which was the most commonly reported symptom among users who self-reported to have tested positive for COVID-19 on Twitter,5 is reported much less frequently for long-COVID (6th). Table 1 presents the relative frequency distributions of 10 commonly-reported long-COVID symptoms from this study and 3 recently-published papers on long-COVID. Fatigue is consistently reported with high frequency across studies, but considerable variations can be observed in the symptom distributions.

Discussion and Conclusion

With limited scientific data available about long-COVID and with medical care globally primarily focusing on treating acute COVID-19 cases, many long-COVID sufferers are turning to social media to discuss their persistent symptoms, finding other sufferers with similar symptoms, and identifying potential solutions to their debilitating symptoms and improving quality of life. The spectrum of symptoms identified from Reddit may provide early insights about long-COVID. Social media data, including from Reddit and Twitter, may help the medical community to better understand long-COVID from the perspective of patients and thus enable them to improve the care they provide. Considering the large volume of data that is generated in social media about this topic, it is necessary to develop automated methods involving NLP for long-term surveillance.

References

Impact of a Clinical Decision Support Tool for Cardiovascular Preventive Care in Community Health Centers: Randomized Trial Results

Rachel Gold, PhD, MPH1,2; Patrick O’Connor, MD, MA, MPH1; Annie Larson, PhD2; JoAnn Sperl-Hillen, MD3; Dave Boston, MD2; Lauren Crain, PhD3; Christina Sheppler, PhD1; Mary Middendorf, BS2; Ann Romer, MS2; John Heintzman, MD2,4; Deepika Appana, BS3

1Kaiser Permanente NW Center for Health Research, Portland, OR; 2OCHIN, Inc., Portland, OR; 3HealthPartners Institute, Bloomington, MN; 4Oregon Health and Science University, Portland, OR

Introduction: Modifiable cardiovascular disease (CVD) risks can be effectively managed, but CVD risk control remains suboptimal in socioeconomically vulnerable populations.1 Primary care providers must consider multiple factors impacting a patient’s CVD risk, and assess which to address for optimal impact, within a brief encounter.2 Electronic health record (EHR)-based clinical decision support systems (CDSS) have been shown to support this process and improve CVD risk management.3–5 However, little relevant evidence on the adoption and impact of such informatics tools comes from health care settings serving low-income patients, such as community health centers (CHCs). Such settings primarily serve patients regardless of insurance coverage. They are rarely resourced to develop their own CDSS. Thus, implementing CDSS proven effective in other settings might efficiently enhance CVD risk management in CHCs. To test this approach, we conducted a randomized trial in CHCs of a non-proprietary, web-based CDSS developed at an integrated care system,6 called CV Wizard©. It processes EHR data through algorithms that reflect evidence-based CVD care and are updated as guidelines evolve. This is one of the first trials evaluating the impact of an effective CDSS developed in an integrated care setting, when implemented in primary care CHCs.

Methods: Design: Cluster-randomized trial; randomization was by CHC organization. As adoption rates necessarily drive the CDSS’ impact, we assessed a) population-level impact in an intent-to-treat analysis (ITT), and b) impact when the CDSS was used, in an effect-of-treatment-on-the-treated (ETOT, or ‘per protocol’) analysis. Setting: 67 clinics, managed by 15 CHC organizations, that share an EHR as members of OCHIN, Inc. The CDSS was activated in the intervention CHCs in September 2018. The presentation will include further details about the intervention’s design and content. Patients: Adults aged 40-75 with high modifiable CVD risk (either diabetes or heart disease and ≥1 uncontrolled CVD risk factor, or an absolute 10% reversible 10-year CVD risk). The CDSS was used at 35% of eligible patients (7.4% of all eligible encounters), and just viewed at 3%; 61% of eligible patients were exposed to the tool at least once. In ITT analyses of clinic patients (n=10,984 intervention and n=7,247 control), there was no significant overall effect on CVD risk. However, intervention clinic patients’ modifiable CVD risk improved 8.4%, significantly more than among control clinic patients (7.4%), among those in the highest risk quartile at baseline (p=.046; baseline reversible risk quartile parameters were as follows: Q1 ≥10–<12, Q2 >12–<15, Q3 >15–<20, Q4 >20). In ETOT analyses, intervention
clinic patients were categorized by how often CV Wizard was used in their index and follow-up encounters (never, once, or more than once) and matched to control clinic patients on demographic and clinical characteristics. Among those for whom the tool was used more than once, intervention patients’ CVD risk decreased significantly more (by an absolute 1.68%) than control patients, whose risk declined by just 0.13% ($p<.001$). Among those in the highest quartile of baseline CVD risk, intervention clinic patients’ CVD risk improved by 4.6%, significantly more than among control patients (2.5%; $p<.001$). Those for whom the tool was used more than once had a statistically, but likely not clinically, significant improvement in diastolic BP (annual change of -3.2 mmHg) versus their matched controls (-2.5 mmHg; $p=.013$) but no change in systolic BP. No other outcomes differed significantly between comparison groups. The presentation will include details about tool adoption rates and user perceptions of the tool’s utility and lessons learned from this study.

**Discussion:** This cluster-randomized trial of a CDSS designed to reduce modifiable CVD risk found few improvements in the main outcome measure when comparing all eligible patients in intervention versus control clinics, after accounting for baseline covariates. However, significant impact was seen among those in the highest quartile of baseline CVD risk in both ITT and ETOT analyses. Adoption of the tool was far lower than that seen in its integrated healthcare setting, where its use was part of standardized workflows; such standardization is not pragmatic across diverse CHCs. Analyses are ongoing to better understand these adoption rates and implementation barriers. More research is also needed to assess the benefits of preferentially targeting higher CVD risk patients who are most likely to benefit. These results may inform future efforts to improve CVD care in CHCs by implementing CDSS and may provide a positive template for implementation of cutting-edge informatics tools in CHCs. The relatively low adoption rates of the CDSS in the CHC setting suggest that obstacles to routine use of such technologies must be addressed for this strategy to improve clinical care outcomes for millions of socioeconomically vulnerable high-CVD risk patients nationwide. In conclusion, informatics tools such as CDSS can improve CVD risk among socioeconomically vulnerable patients; further research is needed to understand which patients benefit most and how to improve tool adoption in the CHC setting.

**References**

Patient Perspectives on an Interactive mHealth Intervention to Improve Symptom Management and Medication Adherence in Early-Stage Breast Cancer: Evidence from the THRIVE Study

Ilana Graetz PhD¹, Rebecca Krukowski PhD², Janeane N Anderson PhD², Mehmet Kocak PhD², Teresa Waters PhD³, Edward Stepanski PhD⁴, Andrea N Curry PhD⁵, Andrew Robles MA¹, Gregory A. Vidal MD PhD²,³, Lee Schwartzberg MD⁵

¹Emory Rollins School of Public Health, Atlanta, GA, USA; ²The University of Tennessee Health Science Center, Memphis, TN; ³University of Kentucky College of Public Health, Lexington, KY; ⁴ConcertAI, Boston, MA; ⁵The West Cancer Center, Memphis, TN

Abstract
The THRIVE study is an ongoing randomized controlled trial of an interactive mHealth intervention with tailored messaging to improve symptom management and medication adherence for women with early-stage breast cancer. Initial evidence from patients demonstrates high level of satisfaction with the interactive app-based intervention and response from oncology team. Tailored text messages provided through the intervention was endorsed as helpful by almost all participants.

Introduction
Despite the efficacy of oral adjuvant endocrine therapy (AET) for improving survival among women with hormone receptor-positive, early-stage breast cancer, adherence rates remain low.¹ AET-related adverse symptoms have been identified as a key reason for non-adherence.¹⁻⁴ The THRIVE study is an ongoing randomized controlled trial of an interactive mHealth intervention with tailored messaging to improve symptom management and medication adherence for women with early-stage breast cancer.⁵⁻⁶ The THRIVE trial is among the first to test the efficacy of an interactive and tailored app-based intervention on AET adherence. We used early evidence from THRIVE intervention participants who completed the 6-month intervention to describe user satisfaction with the app-based intervention.

Methods
We are recruiting and randomizing 300 patients initiating AET to one of three arms: 1) an “App” group (N=100) that receives weekly reminder text messages to use the study app with clinical follow-up for concerning symptoms or non-adherence; 2) an “App+Feedback” group (N=100) that receives weekly reminders to use the app and additional tailored text messages based on their use of the app and reported preferences; or 3) a “Usual Care” group (N=100) that receives usual care only. This analysis presents preliminary results on user satisfaction with the app-based intervention. Intervention participants reported preferences on frequency of text messages, perceived impact of app use on their relationship with their oncology team, and overall satisfaction with the app. App+Feedback participants reported feedback on the tailored text messages, including rating each type of message (medication reminders, symptom management, healthy lifestyle, patient-provider communication, and positive affirmations) from most to least helpful.

Results
Overall, 106 App and App+Feeback participants completed the 6-month intervention by January 2021; 51.9% of patients were ages 60 or older (range 34-82 years), 32.1% were Black, 23.6% had incomes below 200% of the federal poverty level and 52.8% had a 4-year college degree or higher education. Most (80.6%) expressed satisfaction with the app-based intervention. Among 59 participants (55.7%) who reported a symptom or non-adherence event that triggered a follow-up phone call from their oncology team during, 69.5% reported that the response met their expectations and 28.8% that it exceeded their expectations. Most participants (79.3%) reported wanting to receive text messages once per week, 16.0% every two weeks or monthly, and 3.8% daily. Among the 53
participants (50%) randomized to ‘App+Feedback’ who received additional tailored text messages, 66% reported that the messages were “helpful” and 30.2% “somewhat helpful.” Over half (52.9%) rated messages with medication reminders as “very” or “most helpful” 32.1% (Figure 1).

**Figure 1. Tailored Text Message Helpfullness Rating by Topic**

<table>
<thead>
<tr>
<th>Topic</th>
<th>Most Helpful</th>
<th>Very</th>
<th>Helpful</th>
<th>Somewhat</th>
<th>Least Helpful</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medication reminders</td>
<td>⬜</td>
<td>⬜</td>
<td></td>
<td>⬜</td>
<td>⬜</td>
</tr>
<tr>
<td>Symptom management</td>
<td>⬜</td>
<td>⬜</td>
<td></td>
<td>⬜</td>
<td>⬜</td>
</tr>
<tr>
<td>Healthy lifestyle</td>
<td>⬜</td>
<td>⬜</td>
<td>⬜</td>
<td>⬜</td>
<td>⬜</td>
</tr>
<tr>
<td>Patient-provider communication</td>
<td>⬜</td>
<td>⬜</td>
<td>⬜</td>
<td>⬜</td>
<td>⬜</td>
</tr>
<tr>
<td>Positive affirmations</td>
<td>⬜</td>
<td>⬜</td>
<td>⬜</td>
<td>⬜</td>
<td>⬜</td>
</tr>
</tbody>
</table>

% of App+Feedback Participants (N=53)

**Conclusions**

Initial evidence from the THRIVE trial using a mobile app to monitor patients receiving AET demonstrates high level of satisfaction with the interactive app-based intervention and response from oncology team. Tailored text messages provided through the intervention was endorsed as helpful by almost all participants.

**References**

Evaluation of Token Collections and Matching Models to Support Privacy-Preserving Record Linkage (PPRL)

Shaun J. Grannis, MD, MS\textsuperscript{1,2}, Abel Kho MD, MS, FACMI\textsuperscript{3}, Jasmin Phua\textsuperscript{4}, PhD, Suranga N. Kasthuriratne, PhD\textsuperscript{1,2}

\textsuperscript{1}Regenstrief Institute, Indianapolis, IN, USA; \textsuperscript{2}Indiana University, Indianapolis, IN, USA; \textsuperscript{3}Northwestern University, IL, USA, \textsuperscript{4}Datavant Inc.

**Introduction.** Patient information is fragmented across the US health care system in numerous independent clinical and claims repositories with no single unique identifier. Fragmented information hinders public health reporting, and reduces the utility of electronic patient information for clinical research, among other use cases requiring integrated data. Data collaborators increasingly seek to integrate data in a privacy preserving fashion for various reasons, including a lack of a business associate relationship, which limits sharing protected health information. Further, local legislation, business processes, and sociopolitical expectations also limit the exchange of identified health data. Consequently, there has been a steady growth in the use of privacy-preserving record linkage (PPRL) methods, and evidence-based approaches for optimizing PPRL accuracy are needed. One such PPRL use case involves the National COVID Cohort Collaborative (N3C) initiative\textsuperscript{[1]}, which seeks to integrate observational clinical data, including medications, lab results, and procedure data for a COVID-19 cohort from sites across the US.

PPRL methods typically use several token definitions (or simply tokens), which represent combinations of demographic and identifier fields that define a match between two records. For example, one token may be defined as agreement on \{last name\}_\{first initial of first name\}_\{gender\}_\{birth date\}, while another may be defined as agreement on \{Social Security number\}_\{first name\}. Tokens can serve as inputs to a variety of matching models. While the body of PPRL literature is growing, there remains a paucity of evidence-based, best-practice guidelines comparing the performance of various matching models using real-world data. This study seeks to evaluate match performance for three different matching models and for two collections of pre-selected PPRL tokens under consideration for the N3C initiative.

**Materials and methods.** **Token Collections.** Our team has over two decades combined experience matching billions of clinical and claims records across a wide variety of use cases and datasets. Using that experience, authors SG, AK, and JP identified an initial collection of 35 PPRL tokens from which we further prioritized 18 tokens for evaluation (N3C tokens, see Appendix A). We sought to evaluate both the 18-token subset collection and the complete 35-token collection. **Evaluation Dataset.** We previously developed a manually reviewed 30,000-record gold-standard matching dataset randomly sampled from the Indiana Health Information Exchange, which contains > 47 million patient registration records across more than 100 clinical data sources. The HIE represents a unique in-vivo laboratory to evaluate real-world patient matching methods, and the demographics of the HIE catchment area closely mirror the demographics of the US overall, supporting generalizability of findings. A list of the HIE demographic fields can be found in Appendix B. **Matching Models.** We selected 3 matching models: First, a baseline deterministic model wherein a record-pair is declared a match if one or more of the token definitions is satisfied. Additionally, we selected eXtreme Gradient Boosting (XGBoost) classification\textsuperscript{[2]} and Least Absolute Shrinkage and Selection Operator (LASSO) regression\textsuperscript{[3]} methods for evaluation. For all three models, we treated each token as an independent feature. For the two advanced models, we randomly split the dataset into 80% training and 20% test partitions; we optimized model parameters via hyperparameter optimization and applied to the holdout datasets. **Evaluation Metrics.** We compared the performance of the three models across both token collections using sensitivity (recall), specificity, precision (positive predictive value), and F-scores with 95% confidence intervals.

**Results.** The gold standard dataset contained 7,840 (52.3\%) true matches and 7,160 (47.7\%) non matches. **Precision and Specificity.** (Figure 1) Both XGBoost and LASSO models exhibited superior precision and specificity compared with the deterministic model. The 35-token collection exhibited significantly lower precision and specificity than the N3C token collection using the deterministic model. While not statistically significant, a similar trend favoring the N3C token collection was noted for both the XGBoost and LASSO methods. **Recall.** For both token collections, the XGBoost and deterministic methods showed statistically similar recall performance, and both were superior to the LASSO method for both token collections. **F-Score.** The N3C token collection provided the highest F-scores for the XGBoost and deterministic methods. LASSO-based methods outperformed only the deterministic method using all 35 tokens.

**Discussion.** While simple in approach, the deterministic model combined with the N3C token collection performed well, with all performance metrics exceeding 97\%. Our results suggest this method and token combination is a low complexity, technically feasible PPRL approach. Because the XGBoost model outperformed LASSO for both recall and the related F-Score metrics while showing similar precision and specificity, our results suggest that XGBoost may provide a modest improvement in match precision, at the cost of greater complexity. Future work will validate these findings among other datasets and use cases. We will also explore automated token selection methods.
Conclusion. Applying a deterministic patient matching decision model to a large, heterogeneous clinical data source using 18 PPRL tokens produces performance metrics (precision, recall, specificity, and F-Score) above 97%. Additional performance gains may be achieved by applying a more complex XGBoost classification model.

Appendix A: List of 18 tokens prioritized for evaluation in the N3C system. Each token consists of multiple features selected from the list of features in appendix A.

1. LN + 1st initial of FN + G + DOB
2. LN + FN + G + zip5
3. SSN + FN
4. LN + FN + G + zip5 + YB + MB
5. LN (soundex) + FN (soundex) + G + DOB
6. LN + 1st initial of FN + DOB + Zip3,
7. LN (soundex) + FN (soundex) + DOB + Zip3
8. LN + FN + DOB + Zip5
9. FN + Email
10. LN + FN + DOB + zip3
11. FN + TEL
12. LN + FN + DOB
13. SSN + DOB
14. LN + FN + G + DOB
15. SSN + G + DOB
16. LN + 1st 3 characters of FN + G + DOB + zip3
17. LN + 1st 3 characters of FN + G + DOB
18. LN + 1st 3 characters of FN + G + zip5
(zip3 == first 3 numbers of zipcode; zip5 == first 5 numbers of zipcode)

Appendix B: Demographics and Identifiers included in the HIE dataset:
Social Security number (SSN), last name (LN), first name (FN), middle name (MN), gender (G), month of birth (MB), day of birth (DB), year of birth (YB), full date of birth (DOB), street name, zip code (ZIP), city, state (ST), Cell phone Number (TEL), Email

References
Precision VISSTA Study: mHealth Physical Activity Patterns and Patient-Reported Outcomes in Patients with Inflammatory Bowel Diseases

Ashley C. Griffin, PhD, MSPH1,2, Lucas Mentch, PhD3,
Feng-Chang Lin, PhD1, Arlene E. Chung, MD, MHA, MMCi, FAMIA1,2
1University of North Carolina at Chapel Hill, Chapel Hill, NC; 2Carolina Health Informatics Program, Chapel Hill, NC; 3University of Pittsburgh, Pittsburgh, PA

Introduction

Inflammatory bowel diseases (IBDs), which are comprised of Crohn’s Disease and ulcerative colitis, are chronic intestinal disorders of the gastrointestinal tract. IBDs are characterized by cycles of active and dormant states of inflammatory immune response with symptoms such as abdominal pain. Physical symptoms are frequently accompanied by anxiety, depression, or diminished quality of life and may be worsened by IBD therapies. Physical activity may improve physical and psychosocial symptoms, although there may be variations in inflammation related to dimensions of activity such as duration, frequency, or intensity. Much of this research is based on self-reported data which are typically over-reported. Wearable devices provide objective mobile health (mHealth) data on physical activity (i.e., steps, distance, duration), and provide richer data on various dimensions of activity. Little is known about whether there are activity phenotypes, which are observable behavioral physical activity characteristics, and how these may differ relative to patient-reported outcomes. A more nuanced understanding of activity phenotypes could be meaningful in tailoring lifestyle recommendations to improve outcomes.

Objectives

The objectives of this Precision VISSTA study were to identify clusters of physical activity patterns to generate phenotypes and to evaluate the associations between physical activity and various patient-reported outcomes (disease activity and psychosocial domains) in patients with IBDs.

Methods

Precision VISSTA uses data from the Crohn’s and Colitis Foundation of America (CCFA) Partners study, a U.S. Internet-based cohort of adults (18+) with self-reported IBDs where participants have access to a portal to sync physical activity tracking smartphone applications or wearables through a bring-your-own-device model. Participants also complete biannual questionnaires including Patient-Reported Outcomes Measurement Information System (PROMIS) short forms in the domains of depression, anxiety, pain interference, sleep disturbance, social relationships, and fatigue, and complete disease activity questionnaires specific to type of IBD (short Crohn’s Disease Activity Index-SCDAI or Simple Clinical Colitis Activity Index-SCCAI). PROMIS questionnaires are scored using standardized T-scores with a population mean of 50 (SD=10) where higher scores indicate worse symptoms. SCDAI scores range from 0 to 600, and SCCAI scores range from 0 to 19 with higher scores indicating more severe disease.

Participants were included in this analysis if they were enrolled in the CCFA Partners study between 2011 to 2020 and had at least 50% of non-consecutive or consecutive days of wearable activity data within the 6 weeks prior to completing at least one questionnaire. Participants could be included in multiple time periods if they completed more than one questionnaire (~20% of participants (113/543) who completed a questionnaire but did not have sufficient activity data were excluded). Features included in the analysis were daily average for each day that activity was recorded by a participant: number of steps, duration of moderate-to-vigorous activity (minutes), and distance of activity (miles). Features were then averaged during the 6 weeks prior each questionnaire timepoint per participant.

K-means cluster analysis, which is a simple and widely used algorithm that organizes data into non-hierarchical groups, was used to generate the physical activity groupings. K-means uses an iterative approach to assign each data point to the closest mean (i.e., centroid) within each cluster by minimizing the average sum of squared Euclidean distance. The number of clusters (K) was determined by calculating the sum of squared errors (SSE) for different values of K. The value of K was selected where the change in SSE decreased and was becoming plateau, indicating additional clusters produced little value (“elbow point”). The quality of clustering was evaluated using silhouette coefficients to measure the amount of cohesion and separation within and among the clusters. Ranging from -1 to 1, higher coefficients indicate better clustering. Based on these clusters, we conducted a cross-sectional analysis to assess sociodemographic differences between the clusters and to compare means of disease activity and PROMIS scores.
across clusters using one-way analysis of variance (ANOVA). All analyses were conducted in Python 3.7, and K-means was conducted using the k-means Scikit-learn package.

**Results**

The final analytic sample included 430 participants with 285 (66.3%) having Crohn’s Disease and 145 (33.7%) having ulcerative or indeterminate colitis. Participants were primarily female (74.0%), white (95.2%), and attained at least a college degree (77.9%). On average, age was 42 years, BMI was 26, and duration of disease was 15 years. The majority used a Fitbit device (86.3%). Daily on average, participants took 7,893 steps, performed moderate-to-vigorous activity for 41 minutes, traveled 3.5 miles, and burned 520 calories during exercise activities.

K-means cluster analysis identified three clusters. Data within each cluster were assessed to determine a label that represented each cluster’s activity attributes. We labeled the clusters in a way that is interpretable: low, moderate, and high physical activity phenotypes (Table 1). Participants could be in multiple clusters if they completed ≥1 questionnaire and had sufficient activity data. For the 430 participants, there were 1,255 total 6-week periods, of which 423 (33.7%) were classified as low activity, 577 (46.0%) as moderate activity, and 255 (20.3%) as high. Overall, clusters were moderately defined with an average silhouette coefficient = 0.54. Across all patient-reported outcomes (PROs), patients in the low activity cluster had the worst scores and those in the high activity cluster had the best scores. Scores varied significantly across clusters for depression, pain, fatigue, sleep disturbance, social relationships, and short Crohn’s Disease Activity Index (p<0.01). Sociodemographic characteristics varied across clusters with those with low activity being older, having higher BMIs, and longer disease duration compared to the other clusters (p<0.05).

**Table 1.** Patient-reported outcome scores across physical activity clusters.

<table>
<thead>
<tr>
<th>Patient-Reported Outcome Measures, Mean (SD)</th>
<th>Low Activity n=423</th>
<th>Moderate Activity n=577</th>
<th>High Activity n=255</th>
</tr>
</thead>
<tbody>
<tr>
<td>Short Crohn’s Disease Activity Index Score</td>
<td>133.3 (80.3)**</td>
<td>117.0 (71.9)**</td>
<td>102.2 (59.7)**</td>
</tr>
<tr>
<td>Simple Clinical Colitis Activity Index</td>
<td>2.9 (2.1)</td>
<td>2.6 (1.8)</td>
<td>2.3 (2.5)</td>
</tr>
<tr>
<td>PROMIS Anxiety (lower score is better)</td>
<td>50.4 (9.2)</td>
<td>49.3 (8.9)</td>
<td>49.2 (8.9)</td>
</tr>
<tr>
<td>PROMIS Depression (lower score is better)</td>
<td>48.6 (8.3)**</td>
<td>47.1 (7.8)**</td>
<td>46.9 (7.3)**</td>
</tr>
<tr>
<td>PROMIS Pain (lower score is better)</td>
<td>50.7 (9.5)**</td>
<td>48.5 (8.4)**</td>
<td>46.7 (7.6)**</td>
</tr>
<tr>
<td>PROMIS Fatigue (lower score is better)</td>
<td>54.9 (10.7)**</td>
<td>51.5 (10.9)**</td>
<td>50.7 (9.6)**</td>
</tr>
<tr>
<td>PROMIS Sleep Disturbance (lower score is better)</td>
<td>51.2 (7.5)**</td>
<td>49.4 (7.7)**</td>
<td>49.2 (7.8)**</td>
</tr>
<tr>
<td>PROMIS Social Relationships (higher score is better)</td>
<td>50.3 (9.7)**</td>
<td>54.0 (9.1)**</td>
<td>55.3 (9.1)**</td>
</tr>
</tbody>
</table>

* p<0.05, ** p<0.01

The sample was limited to participants in the CCFA Partners cohort who synced their smartphone application or wearable device, which may not be representative of all patients with IBDs. Most consumer-grade wearable devices do not make the details of their algorithms or firmware updates available, so there may be differences in hardware or sensors across brands and devices over the ten-year study period. The clustering approach was limited by the lack of gradient or differentiation between individuals once clusters were established, so it is possible that individuals with similar activity patterns could be in different clusters. The sparse nature of the 6-month longitudinal survey timepoints also precludes causal inferences.

**Conclusion**

Recognition of patterns and changes in lifestyle behaviors by leveraging real-world mHealth data supports opportunities to inform interventions. Patients in the low activity cluster generally reported worse psychosocial and disease activity than those in moderate and high activity clusters. Additional support for physical and psychosocial symptoms and exercise may be valuable for those in low activity subgroups. Future research should examine these findings among more diverse cohorts with more frequent PROs measurements and validate reproducibility.

**References**


1381
Effect of Abbreviation and Acronym Expansion on Patients' Comprehension of their Health Records: A Randomized Trial

Lisa Grossman Liu1*, Meghan Reading Turchioe2*, David Russell3, Annie Myers2, David K. Vawdrey4,1, Ruth M. Masterson Creber2

1Department of Biomedical Informatics, University, New York, NY; 2Department of Population Health Sciences, Division of Health Informatics, Weill Cornell Medicine, New York, NY; 3Department of Sociology, Appalachian State University, Boone, NC, USA; 4Geisinger, Danville, PA

Introduction
Taking effect in April 2021, U.S. federal policy implementing the 21st Century Cures Act will require that all electronic health record information, including clinical notes, be made accessible free of charge to patients.1 An estimated 100 million Americans accessed their own health records online in 2019, compared with 65 million in 2017.2 This transparency has prevented medical errors, increased shared decision-making, and improved health outcomes. Given the positive effects of transparency on health and healthcare, as well as the protections proffered by the 21st Century Cures Act, patient access to records will likely continue to increase.

Unfortunately, medical abbreviations and acronyms hinder patients' comprehension, interpretation, and use of their own health information, especially clinical notes, which may cause miscommunication and diminish the potential value of transparency.3 Abbreviations constitute 30 50% of words in clinical notes, compared to <1% in general text such as news media, which demonstrates the inaccessibility of health records for most U.S. patients. Entire sentences may be composed of abbreviations, for example "50 yo F w/hx BSO pw/ LLQP," translated as "50 year old female with a history of bilateral salpingo-oophorectomy presents with left lower quadrant pain." It is therefore unsurprising that abbreviations cause confusion and other unintended negative consequences, and practical solutions are needed.

The Joint Commission prohibits abbreviations in discharge summaries and other patient-only documents. Unlike patient-only documents, however, clinicians contribute to and use health records such as notes, and prohibiting abbreviations is unrealistic. To address this challenge, the field has focused on post-hoc or automated expansion of abbreviations that are already present, which is a more pragmatic solution. It is undoubtedly important to expand most of these abbreviations before patients will understand them. However, the magnitude of the effect on comprehension has not been quantified. This quantification will be essential to guide policy efforts around expansion. In this two-arm, parallel, individually randomized trial at three urban hospitals, we estimate the effect of expansion on patients' comprehension of abbreviations. To isolate the main effect, we control for patients' prior exposure to the healthcare system, the clinical and written context, and the potential difficulty of comprehension.

Methods
We recruited a purposive sample representative of the U.S. population from 3 medical centers in the New York City area. To control for prior health system exposure and clinical context, only English-speaking adult patients with advanced heart failure (NYHA Class III or IV) currently or recently hospitalized (<1 month ago) were recruited.

Participants received either abbreviations (e.g., "MI") or expansions (e.g., "myocardial infarction"). Six clinicians rated 20 abbreviations and expansions commonly found in advanced heart failure notes as easy, moderate, or difficult for their patients to comprehend. We then chose 10 abbreviations of varied difficulty and their corresponding expansions for inclusion in the trial (Figure 1): [easy] "hrs" (hours), "MD" (medical doctor), "BP" (blood pressure), "ED" (emergency department), [moderate] "yo" (year old), "pt" (patient), "HF" (heart failure), [difficult] "hx" (history), "HTN" (hypertension), "MI" (myocardial infarction).

A coordinator interviewed participants by phone or in-person at the outpatient clinic or inpatient unit. Participants completed a baseline questionnaire to assess demographic and socioeconomic characteristics and were screened for inadequate health literacy. Then, participants were assigned groups (1:1 ratio) using a computer-generated block randomization algorithm (size=4) that concealed allocation sequence. We presented abbreviations and expansions in
a short paragraph to contextualize them, similar to the HPI (history of present illness) section of doctors' notes. Data were recorded in REDCap, a HIPAA-compliant research application.

Based on clinicians' predictions, we expected 89% (8.9/10) overall comprehension in the intervention group and 56% (5.6/10) in the control group. We used a two-sided unequal variances t-test at P<0.05 to analyze differences in overall comprehension (primary outcome), defined as the summary (count) score of the total number of abbreviated or expanded terms comprehended. Fisher’s exact tests at P<0.05 were used to examine differences between abbreviated and expanded versions of each individual term. We conducted bivariate analyses to assess whether baseline characteristics differed by group. Nominal variables were compared using Fisher’s exact tests, while ordinal and numerical variables were compared using Wilcoxon rank-sum tests. In a subgroup analysis of control participants, correlation (numerical variables) and t-test or ANOVA (categorical variables) were used to determine relationships between individual baseline characteristics and overall comprehension (summary score).

Results

Of 41 patients approached between February and November 2020, 29 agreed to participate. We randomized 17 to receive abbreviations (control group) and 12 to receive expansions (intervention group). Overall comprehension scores were significantly greater among patients in the intervention group who received expansions (score 9.3/10 [95% CI, 8.9 to 9.7]) than among patients in the control group who received abbreviations (score 6.7/10 [95% CI, 5.5 to 7.9], P<.001). Participants with an older age, lower educational attainment, and inadequate health literacy comprehended fewer abbreviations.

Discussion

Today, millions more patients access their own health records online than even 5 years ago, and U.S. federal policy will require access beginning April 2021. Unfortunately, abbreviations and acronyms hinder patients' comprehension of their own information, which may cause unintended negative consequences. In this randomized trial, patient comprehension of common abbreviations in health records was poor-to-moderate (67%). However, expanding abbreviations led to an impressive increase in comprehension (93%, P<.001). The findings suggest that post-hoc or automated expansion would substantially improve patient understanding of their own health information. As such, expansion should be encouraged as part of federal policy.

Acknowledgements

We gratefully acknowledge the support and contributions of the patients and healthcare providers who participated in the study. Thank you to George Hripcsak, Suzanne Bakken, and Jessica Ancker for your support.

References


Elham Hatef, MD, MPH, FACPM1, Masoud Rouhizadeh, MS PhD2, Claudia Nau, PhD3, Fagen Xie, PhD3, Ariadna Padilla, MBA3, Lindsay Joe Lyons, BS, LVN3, Christopher Rouillard, MSPH (cand) 4, Mahmoud Abu-Nasser, PhD4, Hadi Kharrazi, MD, PhD1, Jonathan P Weiner1, Douglas Roblin, PhD4

1Johns Hopkins Bloomberg School of Public Health, Baltimore, MD; 2 Johns Hopkins Medical Institute, Baltimore, MD; 3 Kaiser Permanente Southern California, Pasadena, CA; 4 Kaiser Permanente Mid-Atlantic States, Rockville, MD

Introduction

International Classification of Diseases (ICD) coding system have codes for the recording of social needs; however, documentation of non-clinical issues in electronic health records (EHRs) is infrequent compared to medical conditions. ICD codes in EHRs for social needs identification, therefore, may under-report patients with such needs and risks, which makes it difficult for healthcare systems to target “high risk” patients for interventions addressing social needs. Social needs may be discussed with healthcare providers during visits and, therefore, recorded in EHR unstructured data or free-text providers’ notes. These notes might provide a more accurate accounting of social needs; however, traditional approaches for review and abstraction of patient information from the notes are laborious, expensive, and slow. Recent developments in text mining and natural language processing (NLP) of digitized text allow for reliable, low-cost, and rapid extraction of information from EHRs. In this pilot project, we evaluated whether an NLP algorithm could extract valid measures of social needs from Epic-based EHRs in three healthcare systems: Johns Hopkins Health System (JHHS), Kaiser Permanente Mid-Atlantic States (KPMAS), and KP Southern California (KPSCal). The focus of our study was residential instability (i.e., homelessness and housing insecurity).

Methods

The study was conducted independently, in a parallel and coordinated framework across sites. The study population included beneficiaries ≥18 years of age during 2016 through 2019 who received care at JHHS, KPMAS, KPSCal and had at least one provider note in the EHR. To develop hand-crafted linguistic patterns a team of experts at JHHS reviewed ICD-10, CPT, LOINC codes, SNOMED terminologies to identify phrases related to residential instability, and the description of residential instability in public health surveys and instruments (e.g., U.S. Census, American Community Survey). Our team also reviewed phrases derived from a literature review of other studies and the results of a manual annotation process from a past study. To craft the linguistic patterns, the expert team developed a comprehensive list of available codes, specific content areas, and phrases for residential instability, matched them across different coding systems, and developed several phrases and synonyms to describe each content area. Using these phrases each study site included lemma variants of the terms to address variation in describing the residential instability in the providers' notes at their site. Each study site performed pre-processing on the extracted providers' notes, which included (1) cleaning developments and non-word or digital characters, (2) spelling check and correction for mistyped, misspelled, or concatenated words detected during the NLP development process in previous studies, (3) sentence separation, and (4) tokenization (i.e., segmenting text into linguistic units such as words and punctuation). The validation assessment and NLP algorithm logic were identical across sites; however, the “gold standard” for assessment of algorithm validity differed according to data availability: while JHHS and KPMAS used the positive/negative responses to specific residential instability questions in the EHR-based questionnaires, KPSCal used relevant ICD-10 codes and manual annotation of a limited number of patients to develop the gold standard (Table 1).

We applied spaCy’s open-source natural language processor to process and interpret unstructured provider notes. Using the EntityRuler module of the spaCy 2.3 Python toolkit, we created a rule-based NLP system made up of 61 expert-developed patterns that, if present, would represent residential instability. Our patterns included word 'lemmas' and base forms to account for morphological variations (e.g., singular and plural forms) as well as substitutions of different prepositions (e.g., about and for), and synonym words (e.g., house and apartment). We utilized SpaCy’s
PhraseMatcher function to efficiently identify phrases indicating residential instability using the developed patterns. The process included searching each sentence for patterns addressing residential instability and patterns associated with a historical term or a historic date of residential instability. We considered the identification of negated patterns and patterns describing residential instability for someone else or the actual situation (false positive) as the absence of residential instability for the patient. We revised and optimized the patterns through an iterative application of the natural language processor within the training dataset. We completed pattern revision and optimization before model implementation on the validation dataset. Therefore, the validation dataset pattern results did not influence the pattern generation and revision. We calibrated and then validated the algorithm using a split-sample approach. The sample size for the training and validation datasets was selected proportional to the number of available responses to the EHR questionnaires (JHHS and KPMAS) and documented ICD-10 codes and resources available for manual annotation (KPSCal) (Table 1). Validity was assessed at each site by measures of sensitivity and specificity.

Results
Characteristics of the study population and main findings of the study are presented below.

Table 1. Characteristics of the Study Population and Performance of the NLP Algorithm at Each Study Site

<table>
<thead>
<tr>
<th>Study Population (Patient No.)</th>
<th>JHHS</th>
<th>KPMAS</th>
<th>KPScal</th>
</tr>
</thead>
<tbody>
<tr>
<td>~1,200,000</td>
<td>~1,600,000</td>
<td>~4,700,000</td>
<td></td>
</tr>
</tbody>
</table>

NLP Validation

<table>
<thead>
<tr>
<th>Gold Standard</th>
<th>EHR Social Needs Questionnaires</th>
<th>EHR Social Needs Questionnaires</th>
<th>Social Needs ICD-10 Codes, Manual Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Size</td>
<td>1,000 (500+/ 500-)</td>
<td>8,197 (833+,7364-)</td>
<td>300 (150+/150-)</td>
</tr>
<tr>
<td>Patients/ Response No.</td>
<td>1,000 (500+/ 500-)</td>
<td>8,197 (833+,7364-)</td>
<td>300 (150+/150-)</td>
</tr>
<tr>
<td>(with/without residential Instability)</td>
<td>1,000 (500+/ 500-)</td>
<td>8,197 (833+,7364-)</td>
<td>300 (150+/150-)</td>
</tr>
<tr>
<td>Clinical Note No. (an Individual Note)</td>
<td>134,062</td>
<td>78,825</td>
<td>9,575</td>
</tr>
</tbody>
</table>

NLP Algorithm Performance

<table>
<thead>
<tr>
<th>Sensitivity</th>
<th>0.84</th>
<th>0.61</th>
<th>0.96</th>
</tr>
</thead>
<tbody>
<tr>
<td>Specificity</td>
<td>0.96</td>
<td>0.87</td>
<td>0.97</td>
</tr>
</tbody>
</table>

1 Randomly selected from the pool of patients with +/− responses to the residential instability questions in the EHR questionnaires. Randomly selected 50% of the sample to develop the training dataset and reserved the remaining subset for the validation dataset. Total number of available +/− responses to the residential instability questions in the EHR questionnaires. Randomly selected 80% of the total sample to develop the training dataset and 20% for the validation dataset. 2 Randomly selected from patients with documented ICD-10 codes related to residential instability. Randomly split the total sample into 5 subsets (each containing 30 + and 30 - patients for residential instability). Used 4 subsets to develop the training dataset and reserved the remaining subset for the validation dataset.

Conclusion
The consistent performance of this NLP algorithm to identify residential instability in three different healthcare systems suggests the algorithm is generalizable. The consistent and relatively high sensitivity and specificity demonstrate the algorithm’s validity. The variation in the performances among different health systems could be due to different approaches for development of the “gold standard” for assessment of algorithm validity. The development of generalizable NLP algorithms with promising performance will enhance the value of EHRs to identify at-risk patients across different health systems, improve patient care and outcomes, and mitigate socioeconomic disparities across individuals and communities. The NLP algorithm could be further developed to address other social needs and expand internationally.

References
A Hybrid Approach to Semi-Automate the Screening Process for Living Systematic Reviews and Meta-Analysis

Huan He, PhD 1*, Irbaz Bin Riaz, MD, MS 2,4*, Syed Arsalan Ahmed Naqvi, MBBS 2
Rabia Siddiqi, MBBS 3, Nourreen Asghar, MBBS 3
M. Hassan Murad, MD, MPH 4, Hongfang Liu, PhD 1
1Department of AI and Informatics Research, Mayo Clinic, Rochester, MN, USA
2 Department of Oncology, Mayo Clinic, Phoenix, AZ, USA
3 Dow University of Health Sciences, Karachi, Pakistan
4 Mayo Clinic Evidence Based Practice Center, Mayo Clinic, Rochester, MN, USA

Abstract

Systematic Reviews (SRs) and meta-analyses (MAs) are tools to synthesize evidence for important clinical topics. They are frequently used by decision makers as they provide precise estimates of effect for clinically important benefits and harms outcomes with associated certainty of evidence. However, when the field is rapidly evolving, laborious process of systematic reviews cannot keep pace with new evidence and SRMAs are quickly outdated. Thus, it is important to keep the systematic reviews “living”. For truly living systematic reviews (LSRs), the most laborious step of screening thousands of citations to identify few relevant studies must be automated. Thus, we propose a hybrid approach that integrates interactive web-based user interface and multiple natural language processing and machine learning based techniques to screen publication to maintain LSRs.

Introduction and Background

Systematic review is a synthesis technology that is widely used by researchers across many specialties 1. Developing an SR generally requires screening of thousands of citations to identify a small number of studies which meet the inclusion criteria. To maintain an LSR, this becomes a long-term iterative process where new studies must be screened at a regular interval to identify new citations. Several machine learning (ML) and natural language processing (NLP) based methods have been proposed to help with the screening process 2. Those methods could provide effective support and reduce the workload in one screening instance.

However, keeping a SR living requires repetitive screening, for example on a weekly basis, and thus applying these models necessitate users to learn not only the computer programming languages to use the packages, but also the modules and workflow to processing the massive data. Thus, maintaining LSRs continues to be an intensely laborious manual process. To semi-automate the screening process, we present a hybrid approach that integrates interactive web-based user interface and multiple NLP and ML based techniques to iteratively screen incoming citations and thus making the screening process efficient for LSRs.

Method

Studies or citations included in LSR are first identified using a conventional search strategy followed by an automated search strategy. Conventional search strategy is executed in multiple databases (e.g., PubMed, EMBASE, etc.) with customized keywords to retrieve the most recent literatures for a specific clinical question. At the first stage, our platform allows users to either complete the initial screening process utilizing the user-friendly interface or the results from the projects where screening has been completed can be simply uploaded (Fig.1 (B)). The main output of the first stage is the decisions for every record (i.e., included or excluded), which could be used as the input for the next stage. Since our team has conducted systematic reviews on several clinical research projects and accumulated relevant datasets, we could proceed directly to the next stage based on these datasets for each project. At the second stage, the search strategy developed in the first stage is re-used periodically based on the speed of influx for new evidence (e.g., weekly, bi-weekly, or monthly) to retrieve new records and expand the project dataset continuously.

In each small batch update, the pipeline works on the new emerging studies based on several different techniques in three steps (Fig.1(A)). We developed two modes to get new records, namely pull mode and push mode. In the pull mode, our system actively retrieves new records for each project by sending the pre-defined query requests to data sources based on the same search strategy designed in the first stage to make sure all the new coming studies meet the same criteria. The pull mode could be performed automatically at regular intervals or manually per user request. Once

* Those authors have contributed equally as co-first authors
the new studies are pulled from the data source, our system will remove the duplicates. Then, all of the new studies are saved in the project database and set to an unscreened status. In contrast to the pull mode, the push mode passively receives new records from the data sources by using the email subscription services provided by the data sources (e.g., Ovid AutoAlert). Our system checks the email inbox regularly and generates a new study list by parsing new emails’ attachments. Then, similar to the pull mode, each study will be further processed, and the new studies will be saved.

In the second NLP and ML based labeling step, various models based on NLP and ML methods are used to add labels to new studies for further screening. For example, the randomized controlled trial detect model based on support vector machine and convolutional neural network is integrated to detect whether a study is an RCT or not; the national clinical trial (NCT) number detect model based on regular expression extracts the NCT number from study abstract; publication type model detects whether a study is original or follow-up. Those automatically generated labels will be stored in the project database as part of the meta-data and indexed by integrated search engine.

In the third interactive screening step, we present the labeled new studies through an interactive web-based user interface to assist users screening (Fig. 1(B)). In this web-based user interface, all studies and their meta-data are listed by the study status (e.g., unscreened, excluded, included, etc.). At this stage, we facilitate the screening process by an interactive design to reduce workload and improve screening efficiency. For example, the labels generated in the second step for study type (RCT or not) are displayed intuitively with each study to help users identify the study type. Furthermore, the search engine allows for filtering studies by text and the color-encoded inclusion/exclusion keywords highlighting eases the screening process. After the third step, the screened new studies have been saved in the project database with the decisions. The meta-data from already screened citations will be used to train and fine-tune the NLP model which will automate the screening process for LSR in future.

Discussion

We iteratively designed and developed this screening system while maintaining living systematic review projects for first line treatment of metastatic kidney cancer, cancer associated thrombosis, and toxicity of immune checkpoint inhibitors. Qualitative feedback from experienced systematic reviewers suggests that our approach decreases the workload and increases efficiency for creating LSRs. The next steps include ongoing enhancements to the visual interface, implementing fine-tuned NLP models for each LSR based on labeled meta-data generated by initial screening of each project, and formal user testing.

References

Investigating the Narratives of Anti-Asian Hate Speech on Twitter During the COVID-19 Pandemic

Lu He¹, Ya Cheng², Tianyang Zhou², Yongxu Xian²
¹Department of Informatics, University of California, Irvine, Irvine, CA, USA,
²Department of Computer Science, University of California, Irvine, Irvine, CA, USA

Introduction
The COVID-19 pandemic has caused significant fatalities, economic loss, and emotional disturbances around the world. The pandemic is characterized by great uncertainties and controversies in its origin, severity, transmission mechanisms, and coping strategies. Facing great uncertainty and emotional stress, people turned to social media such as Twitter to gather information and express their opinions. Among social media discussions related to COVID-19, hate speeches targeting specific racial groups also arose and increased discrimination against certain racial groups. While recent studies investigated the detection, dissemination patterns, network structures, and associations with demographic features of anti-Asian hate speeches on Twitter,¹ ² ³ few studies have investigated the narratives of hate speech during the pandemic to understand how their creators narrated their arguments to attack and discriminate people of specific racial groups. Understanding the narratives and discourses of hate speech during the pandemic could provide implications into design public health communication strategies to reduce hate speech and associated racism.

In this preliminary, exploratory study, we investigated the topics commonly discussed in hate speech targeting Asian and Chinese populations during the COVID-19 pandemic on Twitter. We found that anti-Asian hate speech on Twitter during the COVID-19 pandemic has two major categories: 1. Attributing the virus origin to a specific country by drawing on conspiracy theories; 2. Calling for political actions to punish the proposed origin country.

Methods
Data collection and preprocessing. We collected Twitter data by choosing keyword queries from Hashtagify.me, a hashtag search engine that allows users to find relevant and commonly used hashtags. We started with a few seed hashtags (#ChinaVirus and #WuhanVirus) and used Hashtagify to identify other hashtags that were commonly used together. We also manually looked through tweets that used those hashtags to make sure that they represented content related to hate speech targeting Asian and Chinese populations. In the end, we used the following hashtags to retrieve tweets: ChinaVirus, WuhanVirus, ChinaLiedAndPeopleDied, ChinaLiedPeopleDie, ChinaMustPay, MakeChinaPay, ChineseBioterrorism, ChineseVirus19, ChineseVirusCorona, ChineseWuhanVirus, XiVirus. Data were collected from January 20th, 2020 to April 12th, 2020. After retrieving tweets containing the identified hashtags, we preprocessed the data to remove image and video links. We further excluded non-English tweets and then removed all common English stop words. Then, we tokenized each tweet’s text and prepared the data for analysis.

Data analysis. To extract the commonly discussed topics, we applied Latent Dirichlet Allocation (LDA), which is an unsupervised probabilistic text mining technique widely used in previous studies.⁴ ⁵ We tested different numbers of topics until the resulting topics were distinguishable and comprehensible. We manually read tweets that contained the topical words returned by LDA to derive the topic names and descriptions. We also calculated the Term Frequency-Inverse Document Frequency (TF-IDF) and identified the most important bigrams and trigrams to enhance the readability of the topics when the topical unigrams were not informative enough to derive the topics.

Results
Descriptive statistics. Overall, 151,315 tweets from 45,529 users were collected from the three-month period. Similar to previous studies, we observed a sharp increase in the tweet volume after President Trump tweeted using the phrase “China Virus” on March 16th, 2020.²

Commonly discussed topics. The three commonly discussed topics are presented in Table 1 below. All example tweets were paraphrased to protect the privacy of the original posters.

<table>
<thead>
<tr>
<th>Topic</th>
<th>High-frequency words and phrases</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Discussion and Conclusion

Through exploratory and unsupervised topic modeling on a large Twitter dataset, we identified two major categories of anti-Asian hate speech during the initial stage of the pandemic. The first category (Topic 1) attributed the origin of the virus to a certain country and attacked Asian and Chinese populations based on this virus origin. Tweets within this category also drew on conspiracy theories (e.g., the country of the proposed virus origin intentionally hid the outbreak) to support their arguments. The second category (Topic 2-3) called for political actions, such as voting to support political leaders/parties that will punish the proposed origin country and boycotting products and businesses in the proposed origin country.

Our preliminary study provides initial insights into how creators of hate speech narrated their posts to attack people of certain racial groups during the pandemic. Public health communication should target these hate speech narratives to clarify virus origins and debunk conspiracy theories used to support their arguments. In addition, we note that hate speech in a public health crisis is not only about the disease itself (i.e., virus origin) but also highly politicized by tying political actions such as voting for specific political leaders/parties and calling for sanctions to the disease.

Our study has several limitations. First, we adopted an unsupervised text mining technique that may have degraded performance on informal and short texts such as tweets. In future work, we plan to develop more advanced machine learning models based on qualitative analysis results to obtain more accurate and in-depth categories of hate speech. Second, our analysis only included data from the initial stage of the pandemic and did not provide insights into how anti-Asian hate speech evolved over time in the later stage of the pandemic. In future work, we will include more tweets to date to analyze the temporal changes of anti-Asian hate speech and how they correspond to different events.

References

Apps for Covid-19 in Germany: Assessment Using the Mobile Application Rating Scale

Felix Holl, M.Sc.¹,²,³, Fabian Flemisch¹, Walter J. Swoboda, M.D.¹, Johannes Schobel, Ph.D.¹
¹DigiHealth Institute, Neu-Ulm University of Applied Sciences, Neu-Ulm, GERMANY; ²IBE, Ludwig Maximilian University of Munich, Munich, GERMANY; ³Institute for Global Health Sciences, University of California, San Francisco, San Francisco, CA

Introduction

Fueled by the wide availability of cellular networks and mobile phones, mobile health (mHealth) applications have become widely used to expand access to care and are being used to monitor communicable diseases.¹² The Covid-19 pandemic, which started in 2019, has put healthcare and public health systems around the globe under immense pressure. mHealth applications have been identified as an essential mechanism to control the spread of Covid-19, and a multitude of apps have been developed.³⁴ This study aims to give an overview of apps for Covid-19 in Germany and assess these apps using the German Mobile Application Rating Scale (MARS-G)⁵.

Methods

Both the Apple App Store and the Google Play Store were systematically searched to identify apps for Covid-19 using the regional setting Germany and the following search terms: “Corona, Corona Warn App deutsch, Corona App, Robert Koch-Institut, Kontaktdaten, Kontaktverfolgen, Covid-19 App, Covid-19 App deutsch, Covid-19”. Only apps available for both operating systems (OS) were included in the analysis to reduce potential bias because of the price differences of smartphones running the different OS. Apps included had to be available in Germany and the German language. In addition, apps meeting the following criteria were excluded: 1) city or county-specific, 2) apps in a pilot stage, and 3) scheduling apps for vaccinations and Covid testing.

The MARS-G, a translated and validated version of the Mobile Application Rating Scale, was used to assess the applications included in the analysis.⁵⁶ Two trained raters independently evaluated the quality of the included apps, each rater using both a device with Android OS and iOS. Rating differences of 2 and greater were resolved.

Results

The search was performed on June 6, 2021, and yielded 64 (Google Play Store) and 157 (Apple App Store) results. Fifty apps were available in both app stores, and six apps were extracted for analysis based on the inclusion and exclusion criteria. Table 1 shows the apps included in the study and their characteristics. All apps are free of charge and do not offer in-app purchases.

Table 1. Apps included in the analysis with characteristics (*I (Information), Mo (Monitoring), Me (Measurement), T (Tracking Information), E (Education), D (Data collection))

<table>
<thead>
<tr>
<th>App name</th>
<th>Apple App Store Category</th>
<th>Google Play Store Category</th>
<th>Target audience</th>
<th>Developer</th>
<th>Components*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corona-Warn-App</td>
<td>Health &amp; Fitness</td>
<td>Health &amp; Fitness</td>
<td>Citizens</td>
<td>Robert-Koch-Institute</td>
<td>Mo, T, E</td>
</tr>
<tr>
<td>Corona Health App</td>
<td>Health &amp; Fitness</td>
<td>Health &amp; Fitness</td>
<td>Citizens</td>
<td>University Hospital Würzburg</td>
<td>D, Me, Mo, T</td>
</tr>
<tr>
<td>Luca App</td>
<td>Utilities</td>
<td>Tools</td>
<td>Citizens</td>
<td>culture4life GmbH</td>
<td>D, Me, Mo, T</td>
</tr>
<tr>
<td>NINA App</td>
<td>Utilities</td>
<td>News &amp; Magazines</td>
<td>Citizens</td>
<td>Federal Office for Civil Protection</td>
<td>I, E</td>
</tr>
<tr>
<td>SafeVac App</td>
<td>Medical</td>
<td>Medical</td>
<td>Citizens</td>
<td>Paul-Ehrlich-Institute</td>
<td>D, Me</td>
</tr>
<tr>
<td>STIKO-App</td>
<td>Medical</td>
<td>Medical</td>
<td>Healthcare providers</td>
<td>Robert-Koch-Institute</td>
<td>I, E</td>
</tr>
</tbody>
</table>
The overall quality of the six rated apps was good (overall 4.15). The best-rated apps were NINA (4.34) and Corona Health App (4.29). The highest-rated sections were functionality (4.4), aesthetics (4.25), and information (4.25). The lowest-rated area was engagement (3.63). Detailed scores are shown in table 2.

Table 2. MARS-G scores of the assessed applications.

<table>
<thead>
<tr>
<th>App name</th>
<th>Engagement</th>
<th>Functionality</th>
<th>Aesthetics</th>
<th>Information quality</th>
<th>Subjective quality</th>
<th>MARS-G score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corona-Warn-App</td>
<td>3.70</td>
<td>4.50</td>
<td>4.67</td>
<td>3.93</td>
<td>4.15</td>
<td>4.20</td>
</tr>
<tr>
<td></td>
<td>(3.60 – 3.80)</td>
<td>(4.50 – 4.50)</td>
<td>(4.67 – 4.67)</td>
<td>(3.86 – 4.00)</td>
<td>(4.0 – 4.3)</td>
<td>(4.19 – 4.21)</td>
</tr>
<tr>
<td>Corona Health App</td>
<td>3.60</td>
<td>4.63</td>
<td>4.33</td>
<td>3.90</td>
<td>3.63</td>
<td>4.29</td>
</tr>
<tr>
<td></td>
<td>(3.40 – 3.8)</td>
<td>(4.50 – 4.75)</td>
<td>(4.33 – 4.33)</td>
<td>(3.80 – 4.00)</td>
<td>(3.5 – 3.75)</td>
<td>(4.21 – 4.37)</td>
</tr>
<tr>
<td>Luca App</td>
<td>3.80</td>
<td>4.88</td>
<td>4.38</td>
<td>3.63</td>
<td>4.23</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(3.80 – 3.8)</td>
<td>(4.75 – 5.00)</td>
<td>(4.33 – 4.33)</td>
<td>(3.25 – 4.00)</td>
<td>(4.17 – 4.28)</td>
<td></td>
</tr>
<tr>
<td>NINA App</td>
<td>3.90</td>
<td>4.50</td>
<td>4.67</td>
<td>4.25</td>
<td>4.36</td>
<td>4.34</td>
</tr>
<tr>
<td>SafeVac App</td>
<td>3.20</td>
<td>4.25</td>
<td>3.84</td>
<td>4.42</td>
<td>3.75</td>
<td>3.93</td>
</tr>
<tr>
<td></td>
<td>(3.20 – 3.20)</td>
<td>(4.25 – 4.25)</td>
<td>(3.67 – 4.00)</td>
<td>(4.33 – 4.50)</td>
<td>(3.5 – 4.00)</td>
<td>(3.91 – 3.95)</td>
</tr>
<tr>
<td>STIKO-App</td>
<td>3.60</td>
<td>3.63</td>
<td>3.67</td>
<td>4.60</td>
<td>4.25</td>
<td>3.88</td>
</tr>
<tr>
<td></td>
<td>(3.40 – 3.80)</td>
<td>(3.50 – 3.75)</td>
<td>(3.67 – 3.67)</td>
<td>(4.60 – 4.60)</td>
<td>(4.20 – 4.30)</td>
<td>(3.79 – 3.96)</td>
</tr>
</tbody>
</table>

Discussion

The target audience for all but one app is private citizens. The overall quality of the apps is rated high, despite the short development time of the apps. The fact that all but one applications were developed by federal institutions or a public academic institution could explain the high Information quality rating. The low scores in the engagement section show that there is still room for improvement to increase the uptake of Covid-19 apps in Germany. This study is limited by the fact that apps were searched and downloaded on June 6, 2021. Given the dynamic of the Covid-19 situation, apps get updated often. In addition, the Luca app received a high rating but has been heavily criticized for its security vulnerability, which is not reflected in the rating.

Conclusion

To our knowledge, this is the first study that identified and assessed Covid-19 apps available in app stores in Germany. The study has shown that despite the good quality in aspects such as information and functionality, there is still room for improvement in the engagement section.

References

Assessing the Impact of COVID-19 on Clinician Electronic Health Record Use

A Jay Holmgren, PhD MHI1, N Lance Downing, MD,2 Mitchell Tang, BA,3 Christopher Sharp, MD,2 Christopher Longhurst, MD,4 Robert S Huckman, PhD3

1University of California, San Francisco, San Francisco CA; 2Department of Medicine, Stanford University, Stanford, CA; 3Harvard Business School, Boston, MA 4Department of Biomedical Informatics, University of California, San Diego, San Diego, CA

Introduction

The COVID-19 pandemic issued in a sudden and dramatic change to many aspects of the health care delivery system, including clinician workflows and relationship with the electronic health record (EHR). In a matter of weeks, health care organizations shifted ambulatory care delivery to include phone- and video-based telemedicine encounters.(1) Clinicians still delivering face-to-face care were forced to re-structure many aspects of their practice in order ensure the health and safety of their patients and staff. The format, structure, management and delivery of care have been considerably altered in response to the pandemic. For example, during virtual encounters clinicians can no longer collect vitals or conduct the physical exams that inform clinical assessment, reasoning, and decision making. In addition, communication nuances may be lost, and communication with patients and across clinical teams may be more difficult. Taken together, these factors result in greater reliance on the electronic health record (EHR) to access historical data, complete documentation, and engage in communication with patients and other clinicians.

These changes in clinical work may exacerbate existing issues of burnout and cynicism amongst clinicians. Further, if the increased EHR burden is driven by activities such as reviewing patient history or secure messaging with patients, these new demands on clinician time will be largely concentrated in tasks that are currently non-reimbursable in our current billing system. In order to inform policymakers and health systems leaders interested in addressing clinician burnout, working to ensure EHRs support clinicians in times of crises, and developing sustainable payment policy for telemedicine and virtual care, it is critical to examine the short and long-term impacts of COVID-19 on clinician EHR use. To address this pressing need, we used a novel dataset of national, longitudinal EHR metadata to examine the impact of the COVID-19 pandemic on a variety of dimensions of clinician work.

Methods

We aggregated EHR active use metadata from 366 health systems, comprising nearly the entire US customer base of the largest electronic health record vendor, Epic Systems, to measure clinician EHR work. This granular data describes the amount of time a clinician spends actively using the EHR per day, broken down by four functions (Clinical Review, InBasket messaging, Orders, and Notes), as well as the amount of time spent after-hours (“pajama time”) based on clinician schedules. This metadata measures only “active use” time spent working in the EHR, including clicks, mouse movement, and keystrokes, and stops counting if no inputs are detected for five seconds, consistent with previous studies using Epic Systems metadata.(2) We limited our metadata to clinical care EHR use, and excluded time spent in the EHR for other purposes such as research or customization. Our sample included all clinicians with scheduled appointments, including physicians and advance practice practitioners such as physician assistants and nurse practitioners.

We also measured daily volume by number of patient encounters per clinician, as well as number of messages received per day by clinicians, categorized by message source: results, patient, team, prescription, system-generated, and other. All data was then aggregated to the organization-week level to create a longitudinal dataset from December 29, 2019 through December 27, 2020.
We used an ordinary least squares regression with an event study framework to estimate the impact of the pandemic over time on total EHR time per day and after-hours time per day, using the imposition of the California shelter-in-place (SIP) order as our date for the onset of the pandemic, with the two weeks prior as the baseline week. Both models control for daily volume and organization-level fixed effects to address time-invariant omitted variable bias. Next, we calculated descriptive statistics for EHR time per day, broken down by each of the four primary functions, as well as the amount of InBasket messages received per day by source, relative to pre-pandemic baseline levels from 12/29/19 – 3/14/2020. Finally, to estimate the impact of each additional message received on clinician total EHR time per day, we created an ordinary least squares regression model with each message type as the independent variables of interest, with organization and calendar week fixed effects. All models use robust standard errors clustered at the organization-level.

Results

Total EHR time per day and after-hours “pajama time” in the EHR per day decreased significantly in the first 10 weeks of the COVID pandemic compared to the pre-pandemic baseline. From July through December of 2020, however, we found clinicians spent significantly more time in the EHR per day and more time working outside of their scheduled clinic hours (Figure 1). This increase appears to be driven by Clinical Review and InBasket messaging time (Figure 2), while Notes and Orders returned to their pre-pandemic baselines. From April through December 2020, on average clinicians received 52.0% more messages from patients per day compared to pre-pandemic levels. Messages from other care team members also increased by an average of 4.8% from 7/2020–12/2020. In our fixed effects regression, we found each patient message increased total EHR time per day by 2.32 minutes (p<0.01).

Discussion

After a sharp decrease during the onset of the pandemic, US clinicians have spent significantly more time in the EHR each day, as well as increased “pajama time” outside of scheduled clinic hours during the COVID-19 pandemic. Outside of holidays, this increase remained relatively constant through 12/2020, despite fluctuations in COVID case counts, suggesting that changes to workflow to incorporate telemedicine is a more likely driver of the effect than COVID levels. This effect was driven by increases in time spent in functions outside of the traditional patient encounter—Clinical Review and InBasket. Additional time spent on Clinical Review may reflect greater reliance on patient history and documentation from the EHR in the absence of physical exams during telemedicine encounters for organizations providing care virtually. The increase in InBasket time is almost entirely explained by a large increase in messages received from patients, which are by far the most time-consuming message clinicians receive. It may be that the COVID pandemic encouraged substitution from in-person visits to asynchronous communication such as secure messaging. Federal payment policy should take this new demand on clinician time into account when designing sustainable reimbursement policy for the post-COVID era.

References


Ming Huang1, Aditya Khurana2, George Mastorakos2, Andrew Wen1, Huan He1, Liwei Wang1, Sijia Liu1, Yanshan Wang1, Julie E Prigge3, Brian A Costello3, Nilay D Shah1, Henry H Ting1,4, Christi A Patten5,6, Jung-wei Fan1, and Hongfang Liu1*
1Department of AI and Informatics, Mayo Clinic, Rochester, MN, USA; 2Mayo Clinic Alix School of Medicine, Mayo Clinic, Scottsdale, AZ, USA; 3Center for Connected Care, Mayo Clinic, Rochester, MN, USA; 4Department of Cardiovascular Medicine, Mayo Clinic, Rochester, MN, USA; 5Center for Clinical and Translational Science, Community Engagement Program, Mayo Clinic, Rochester, MN, USA; 6Department of Psychiatry and Psychology, Mayo Clinic, Rochester, MN, USA

Introduction: With the advent of Coronavirus Disease 2019 (COVID-19), millions of non-urgent and non-COVID-19 medical encounters were postponed or cancelled by patients and health systems to reduce the risk of COVID-19 infection during in-person visits and prevent the virus spread [1, 2]. For continued healthcare access, most clinic visits have transitioned to online platforms for healthcare access including COVID-19 diagnosis and treatment [3]. Patient portals and other digital platforms hold promise as sustainable and scalable health system intervention strategies to improve access to healthcare for COVID-19 diagnosis and treatment and other healthcare issues [4]. Through the patient portals, patients can receive educational information on COVID-19 preventive care measures, use online triage forms (self-checker and E-visit) for COVID-19 symptom assessment, and send and receive portal messages related to their COVID-19 diagnostic tests and results. If the positively tested patients are at risk, patients can communicate with their providers about COVID-19 care plan. In this study, we propose to characterize patients and their use of asynchronous virtual care for COVID-19 via a retrospective analysis of patient portal messages. The findings can provide insights into the frequency of portal messaging utilization by patients for addressing COVID-19 crisis and impact on patients with respect to COVID-19 related concerns.

Methods: We collected over 2 million portal messages generated by patients between February 1 and September 20, 2020 at Mayo Clinic, a large multi-specialty academic health system. We filtered the patient-generated messages (PGMs) associated with COVID-19 using relevant keywords (e.g., COVID-19, Pandemic, Coronavirus, SARS-CoV-2, and 2019-nCoV) and their synonyms and morphological variations. We then analyzed the distribution of different patient populations by stratifying the unique patients with respect to their personal and social conditions including age, gender, marriage, ethnicity, race, language, and residence and calculated the daily numbers of total PGMs on COVID-19. We summarized reasons for patient utilization of portal messages for accessing COVID-19 related care such as diagnosis, testing, and treatment and seeking supports for various issues including appointment postponement and mental health problems due to COVID-19. We analyzed the PGMs used for assessing COVID-19 symptoms and discussing COVID-19 diagnostic tests and results and care plan to understand the message use for COVID-19 diagnosis and treatment. In addition, we examined other healthcare issues caused by COVID-19 reported in the portal messages to understand COVID-19 impacts on health services and patients.

Results: The majority of PGMs on COVID-19 pertained to COVID-19 symptom self-assessment (42.53%) and COVID-19 tests and results (32.44%) (See Table 1). As shown in Figure 1. The PGMs on COVID-19 symptom self-assessment and COVID-19 tests and results had dynamic patterns (two peaks and three surges) similar to the newly confirmed cases in the US and Minnesota. The trend of PGMs on COVID-19 care plan is in good agreement with that of newly hospitalized cases and deaths. After an initial increase in March, the PGMs on issues such as cancellations and anxiety regarding COVID-19 had a declining trend (See Figure 2). Patients who were 30-64 years old, married, female, white, or urban residents were more likely to send portal messages. The messaging disparity was exacerbated among patients who sent portal messages on COVID-19.

Conclusion: During the COVID-19 pandemic, patients increased portal messaging utilization to address healthcare issues about COVID-19 (particularly, symptom self-assessment and tests and results). The portal messages on COVID-19 could reflect the overall development of COVID-19 and its impact on patients over time. The use disparity of patient populations indicates an opportunity to increase patient engagement in patient portals for minority and rural populations for addressing COVID-19 crisis.
Table 1 Numbers and percentages of PGMs on COVID-19 related care and other healthcare issues

<table>
<thead>
<tr>
<th>Theme</th>
<th>Category</th>
<th>PGMs on COVID-19 (N=403,116)</th>
<th>Count</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>COVID-19 related care</td>
<td>Self-checker</td>
<td></td>
<td>171,438</td>
<td>42.53</td>
</tr>
<tr>
<td></td>
<td>E-visit</td>
<td></td>
<td>4,619</td>
<td>1.15</td>
</tr>
<tr>
<td></td>
<td>Tests &amp; results</td>
<td></td>
<td>130,771</td>
<td>32.44</td>
</tr>
<tr>
<td></td>
<td>Care plan</td>
<td></td>
<td>4,202</td>
<td>1.04</td>
</tr>
<tr>
<td>Other issues caused by the COVID-19 pandemic</td>
<td>General issues</td>
<td></td>
<td>13,793</td>
<td>3.42</td>
</tr>
<tr>
<td></td>
<td>Postponement</td>
<td></td>
<td>29,023</td>
<td>7.20</td>
</tr>
<tr>
<td></td>
<td>Cancellation</td>
<td></td>
<td>20,266</td>
<td>5.03</td>
</tr>
<tr>
<td></td>
<td>Anxiety</td>
<td></td>
<td>22,671</td>
<td>5.62</td>
</tr>
<tr>
<td></td>
<td>Depression</td>
<td></td>
<td>4,039</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>Suicidal ideation</td>
<td></td>
<td>300</td>
<td>0.07</td>
</tr>
</tbody>
</table>

Figure 1 Daily numbers and weekly smoothing averages (WSAs) of PGMs on COVID-19 related care (diagnosis and treatment) – (A) COVID-19 symptom assessment via self-checker, (B) COVID-19 symptom assessment by providers via e-visit, and discussions on COVID-19 tests and results (C) and care plan (D)

Figure 2 Daily numbers and weekly smoothing averages (WSAs) of PGMs on COVID-19 related other healthcare issues – (A) general issues due to COVID-19, (B) postponement, (C) cancellation, (D) anxiety, (E) depression, and (F) suicidal ideation

Reference:
Deep Significance Clustering (DICE) a Heterogeneous Population

Yufang Huang, PhD1; Yifan Liu, MS1; Peter A D Steel, MA, MBBS1; Kelly M. Axsom, MD2; John R. Lee, MD1; Sri Lekha Tummalapalli, MD, MBA, MAS1; Alison Hermann, MD1; Rochelle Joly, MD1; Fei Wang, PhD1; Jyotishman Pathak, PhD1; Lakshminarayanan Subramanian, PhD3; Yiye Zhang, PhD, MS1

1Weill Cornell Medicine, New York, NY; 2Columbia University Medical Center, New York, NY; 3New York University, New York, NY

Introduction

Deep significance clustering (DICE) is an end-to-end machine learning framework to assist with developing risk stratified interventions for a heterogenous population. DICE applies in scenarios where groups of patients may have similar risk level of an outcome but require different interventions due to differing root causes of the risk. In a heterogeneous population, while classification algorithms can produce a predicted risk, the same risk level can be obtained from patients with very different feature manifestations. Thus, predicted risk alone do not provide enough information about how to choose reasonable clinical treatment as the next step. Similarly, regular clustering algorithms are only unsupervised without a focus on the outcome, thus potentially generating clusters that are not clinically meaningful. In these contexts, clearly partitioned patient groups, or precisely predicted individualized risks, without a bridge to the next clinical steps, bring limited translational value. (1) Yet, few existing clustering and classification algorithms jointly achieve outcome-driven risk stratification in an integrated fashion customized to health data. (2) DICE addresses the practical needs in medicine by generating outcome-driven clusters whose members have both similar risk levels and root causes for the outcome, thus supporting risk-level specific intervention development.

Methods

DICE has 3 main components: long short-term memory (LSTM) autoencoder for representation learning, k-means for clustering, and logistic regression for classification (Figure 1). DICE uses a combined objective function and a constraint that requires statistically significant association between the outcome and cluster memberships. The statistical significance is determined using p-values in testing a statistical hypothesis of equal outcome distribution across clusters. The combined objective function and constraint serve to force DICE to learn representations that lead to clusters that are discriminative to the outcome of interest. Furthermore, a neural architecture search (NAS) is designed with an alternative grid search for the number of clusters and hyper-parameters in the representation learning. DICE back-propagates through the cluster membership classification component, to use cluster membership probabilities as input to predict the outcome to ensure that patients in the same cluster have similar risk levels.

Figure 1. The framework of the proposed deep significance clustering (DICE).
We evaluated DICE using 2 datasets with different structure and outcome rates extracted from electronic health records at 2 urban hospitals. Outcomes and patient cohorts evaluated include discharge disposition to home among heart failure (HF) patients and acute kidney injury (AKI) among COVID-19 patients. HF data included adult patients (n=1,585) from years 2014 to 2018 who were treated in the inpatient Medicine service in the hospital; 36.8% of the patients were discharged to home from the hospital. COVID-19 data included adult patients (n=1,002) who were admitted to the hospital in 2020 from the ED; 30.4% of the patients developed AKI subsequently during hospitalization. We compared DICE with baseline methods including (1) PCA for representation learning followed by k-means clustering (PCA (k-means)), (2) AE for representation learning followed by k-means clustering (AE (k-means)), and (3) AE for representation learning with classification followed by k-means clustering (AE w/ class. (k-means)). The number of clusters experimented were set to 2 through 5. The sizes of the representation dimension searched were 20 through 100 for HF and 10 to 20 for COVID-19. Experiments were conducted in PyTorch.

Table 1. Performance evaluation of DICE against baselines.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Baseline</th>
<th>Silhouette</th>
<th>Calinski-Harabasz</th>
<th>Davies-Bouldin</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>HF</td>
<td>PCA (k-means)</td>
<td>0.097</td>
<td>16.1</td>
<td>2.609</td>
<td>0.773±0.061</td>
</tr>
<tr>
<td></td>
<td>AE (k-means)</td>
<td>0.281</td>
<td>68.1</td>
<td>1.744</td>
<td>0.712±0.067</td>
</tr>
<tr>
<td></td>
<td>AE w/ class. (k-means)</td>
<td>0.346</td>
<td>200.0</td>
<td>1.304</td>
<td>0.818±0.058</td>
</tr>
<tr>
<td></td>
<td>DICE</td>
<td>0.484</td>
<td>212.2</td>
<td>0.864</td>
<td>0.834±0.054</td>
</tr>
<tr>
<td>Covid-19</td>
<td>PCA (k-means)</td>
<td>0.188</td>
<td>30.0</td>
<td>1.840</td>
<td>0.738±0.087</td>
</tr>
<tr>
<td></td>
<td>AE (k-means)</td>
<td>0.462</td>
<td>162.8</td>
<td>0.841</td>
<td>0.686±0.091</td>
</tr>
<tr>
<td></td>
<td>AE w/ class. (k-means)</td>
<td>0.266</td>
<td>92.4</td>
<td>1.124</td>
<td>0.734±0.087</td>
</tr>
<tr>
<td></td>
<td>DICE</td>
<td>0.514</td>
<td>253.6</td>
<td>0.664</td>
<td>0.777±0.083</td>
</tr>
</tbody>
</table>

Discussion

DICE was motivated to join concepts of machine learning and statistics as a customized machine learning algorithm specifically for applications in medicine. The trained outcome-driven cluster memberships by DICE can be used to assign new patients into risk-stratified subgroups for specific interventions in a predictive fashion. These features of DICE were demonstrated in the evaluation across data size, variable types, outcome rates, clinical areas, and clinical settings. Beyond HF and COVID-19, DICE may have the potential to be used in other clinical areas to facilitate subtype-specific care and clinical pathways for clinical decision support.

References

Leveraging the MedDRA Biomedical Terminology and Weak Labeling for Mental Health Symptom Surveillance from patient Notes

Marie Humbert-Droz, PhD\(^1\), Suzanne Tamang, PhD\(^2\), Olivier Gevaert, PhD\(^1\)

\(^1\)Stanford Center for Biomedical Informatics Research, Stanford, CA; \(^2\)Stanford University Department of Biomedical Data Science, Stanford, CA

**Introduction**

Syndrome surveillance tools allow for the monitoring and identification of public health indicators of an upcoming outbreak. It uses clinical features before an initial diagnosis is made, allowing for real-time disease surveillance.\(^1\) Although clinical syndrome surveillance systems are well developed to monitor somatic symptoms from patient records, they lack greatly the ability to monitor mental health indicators. While some tools monitoring the general population’s mental health are available, such as national surveys, they are significantly lagged and require substantial time and resources. Both are weaknesses when it comes to syndrome surveillance. Indeed, a syndrome surveillance system needs to be timely, i.e. the data has to become available shortly after collection.\(^1\) Such a system should also require little resources in order to be implemented as easily in developed economies as well as in developing countries. This last point being especially important in the context of monitoring a worldwide pandemic.

Electronic Health Records (EHR) contain myriads of individual level information, and are available shortly after the patient encounter, making it an ideal source of information to build a syndrome surveillance system. Unfortunately, the coded part of EHRs, containing readily available diagnosis information does not necessarily reflect the patient’s mental state or symptomatology. However, such information is contained in free-text form, buried in the encounter notes. Natural Language Processing (NLP) and Information Extraction (IE) methods are great tools for extracting such information from clinical narratives. It is well established that the COVID-19 pandemic and associated social restrictions, involving several levels of lockdown, have had a significant impact on the population’s mental health. Such information as anxiety and depressive state is poorly documented in EHRs as it is not always associated with a definitive diagnosis. However, this information is abundantly present in clinical notes. In this work, we present a year-long analysis of clinical notes from Stanford Healthcare (SHC) during the 2020 pandemic, showing a correlation between the evolution of the pandemic and the rise of anxiety and depressive symptoms, demonstrating that an NLP approach can improve the account for mental health in syndrome surveillance systems.

**Methods**

The data source that was used to conduct our study is the STARR-OMOP dataset. STARR-OMOP is Stanford Electronic Health Record data from its two Hospitals in a Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM), with linked de-identified free-text patient notes. To extract as much clinical information as possible from patient notes, we have designed a hybrid tool, taking advantage of existing NLP libraries of industrial quality.
and medical ontologies to annotate all clinical concepts within MedDRA\textsuperscript{5}. Assertion status (polarity, experiencer and temporality) is determined through a weak labeling approach\textsuperscript{4}.

\textit{Pipeline description}

The MedDRA tagger pipeline is composed of 3 modules: the annotation module, the weak supervision labeling module and the hierarchy mapping module (Figure 1a). To annotate all clinical concepts efficiently, we are taking advantage of the fast implementation of core NLP functions of spacy using a scispacy language model\textsuperscript{3}, as well as building a custom rule-based string matching tool. We use the MedDRA terminology and match the string to the lower case attribute of the text tokens. Each term matched by the tool is assigned its medDRA ID number as label for further analysis. Each concept is extracted with a context window of $n$ tokens before and after the concept to help determine the assertion status. A set of rule-based functions is written and applied to each extracted term and its context window using the efficiency of the Snorkel package\textsuperscript{4}. Each term is labeled for polarity (PRESENT/ABSENT), experiencer (PATIENT/FAMILY) and temporality (PRESENT/HISTORY). Finally, each term is mapped to each level of the MedDRA hierarchy using the hierarchy mapping module to ease further analysis. The MedDRA tagger is then used to extract all clinical terms from the patient notes, for the year 2020 – as well as three years prior to establish a baseline. The 2020 data contains 575,199 patients and 4,550,255 notes. The baseline contains an average of 600,209 patients and 4,125,779 notes. The performance of the MedDRA pipeline will be assessed at the mention level against manual annotations of 100 notes with 10 symptoms concepts (5 somatic and 5 psychiatric). The notes are currently in the annotation phase and the resulting precision, recall and F1 scores will be computed.

\textit{Results}

The processing of all the patient notes for the year 2020 has shown a significant increase in psychiatric disorder mentions starting in March 2020, reaching a peak in April 2020 and reducing to a baseline in June 2020. Psychiatric disorders mentions are higher for the year 2020 than the averaged three years prior (Figure 1b.). Next, Figure 1c shows the breakdown of the psychiatric disorders mentions at the MedDRA preferred term level, shown as the difference between the 2020 mentions count and the baseline, and highlighting anxiety symptoms as the driving source of the increase in psychiatric disorders mentions. Finally, Figure 1d shows the top seven psychiatric mentions from the notes and top seven psychiatric diagnosis codes from the coded part of EHR, highlighting a tenfold signal difference, reinforcing the initial hypothesis that mental health information is better documented in the notes than the coded part of EHR.

\textit{Discussion}

Our hybrid clinical concept extraction tool demonstrated promising performance to extract patient information at scale, with emphasis on mental health disorders. A notable increase in anxiety and depression symptoms at the beginning of the shelter-in-place order in the San Francisco Bay Area in early 2020 can be visually observed, when compared to the period three years prior. Such a procedure to automatically extract mental health data from deidentified patient notes could easily be pooled across disparate institutions and implemented to provide real-time information, thus addressing the gap in the detection of mental health indicators in modern syndrome surveillance systems.

\textit{References}

5. International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH), MedDRA Data Retrieval and Presentation: Points to Consider, 2016
Opportunities to Improve Social Determinants of Health Screening Implementation Through Training and Support for Providers: Implications for Health Information Technology

Bradley E. Iott, MPH, MS1, Jessica Pater, PhD, MS2, Shauna Wagner, BSN, RN2, Tammy Toscos, PhD, MS2, Tiffany Veinot, PhD, MLS1

1University of Michigan, Ann Arbor, Michigan; 2Parkview Research Center, Fort Wayne, Indiana

Introduction

Screening for patients’ individual social needs, also known as the social determinants of health (SDOH) at a larger scale, is becoming increasingly common as American healthcare transitions towards value-based payment models.1 SDOH screening creates opportunities for providers to identify patients’ needs and to improve health outcomes by intervening via tailored care and referrals to social service agencies to address social needs.2 While there is much interest in implementing SDOH screening in healthcare settings, there has been little academic study of screening implementation, including identifying best practices and the impact of screening work on providers, staff, and clinical workflows3,4. Here, we describe barriers and opportunities for providers related to training and support needed for expanded SDOH screening implementation and discuss opportunities for health information technology (HIT) to facilitate social referrals.

Methods

Parkview Health is a non-profit health system featuring nine hospitals in northeast Indiana and northwest Ohio. To identify clinical areas with formal or informal SDOH screening and initial interview respondents, we held two stakeholder meetings in November 2019 with 31 healthcare providers. Using purposeful sampling, we then conducted semi-structured qualitative telephone interviews with 36 providers, staff, and administrators at Parkview. Interviews were recorded, professionally transcribed, and, using an iterative process, two coders applied structural, in vivo, and open codes (Cohen’s Kappa = 0.637). Second-round coding resulted in the creation of themes based on initial coding.

Results

Many providers (n = 15, 41.7%) requested additional training and support related to SDOH screening and response due to limited coverage in their formal training. Moreover, interviewees indicated that SDOH work has been impacting a greater number of provider and staff roles than the traditional focus primarily on social workers, which may not have formal training, indicating the need for accessible training and support during SDOH screening implementation for a wide range of provider and staff roles (Table 1). As part of this, it is important for there to be clear communication as to each provider’s roles and responsibilities concerning SDOH screening, disclosure, and follow-up in the health system. Furthermore, providers described the need for support in using SDOH data stored in the EHR, which may be difficult to access, as one administrator explained, “it is for [others ’] access and nobody else is seeing it... the way we’re all using it is so personal... that... people aren’t looking for it or knowing what to look for.”

Providers suggested the potential need for multi-modal education about the workflow of SDOH screening, including emails, huddles, education sessions, and the opportunity for staff to provide feedback and suggestions. One administrator running a program that provides medications to patients at a reduced cost described how an email notification of changes to a referral process had been ineffective in the past and advocated individual and phone training instead (Table 1). Similarly, an administrator from a cardiology group described the need for SDOH education sessions for staff and soliciting staff input on SDOH screening workflow.

Providers described a need to know what next steps to take following a patient’s disclosure of a social need (Table 1). Multiple providers expressed concern with not knowing how to help a patient with their social needs, as an administrator from the Women’s and Children’s service line describes, “if somebody tells me they’re homeless, what am I supposed to do with that information without having really a lot of support... it’s a lot easier to get people to screen if they know that they’re not going to be up at night, like ‘oh my gosh, somebody’s homeless and I don’t know what to do with them.’” Knowing who to hand patients off to was perceived as an important facilitator of referrals.
Table 1. Themes and Representative Quotes.

<table>
<thead>
<tr>
<th>Theme</th>
<th>Representative Quotes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training to conduct SDOH screening is needed</td>
<td>“[SDOH screening] is a complicated thing and... not all healthcare providers and staff are trained to know what to do to address [SDOH]” (Medication Assistance Program Administrator)</td>
</tr>
<tr>
<td></td>
<td>“It’s long been the domain of the social worker and now nurses and navigators and social support specialists are being asked to do this work and providers are being asked to look at this data and if it wasn’t part of their original training and education, then it might feel like, ‘not my responsibility’ or ‘outside of my wheelhouse’” (ER &amp; ICU Educator)</td>
</tr>
<tr>
<td>Need for a “multi-modality education” about SDOH workflow</td>
<td>“There was a communication breakdown... It was just an email that went out to everybody that said, ‘hey this is changing’ and some of the offices would have benefitted from one-on-one training or even a phone call... that walked them through the changes and what was needed... just helping them understand the importance and how it works” (Medication Assistance Program Administrator)</td>
</tr>
<tr>
<td>Need for resources and training to support referrals</td>
<td>“it’s not just about collecting that data but making sure that everyone knows if there is a concern, here’s what I need to do next for this patient.” (Pharmacy Administrator)</td>
</tr>
</tbody>
</table>

Discussion

Given the potential for referrals by providers to social care services to reduce social needs and improve patient health outcomes, there is growing interest in standardizing and expanding SDOH screening in healthcare settings. We show that healthcare teams will likely require training and ongoing support to successfully screen and refer patients to relevant resources. Given the importance of facilitating referrals both within the healthcare system and to outside organizations, community resource referral platforms, a form of HIT featuring social services agency directory databases, electronic referrals, outcome tracking, and electronic health record integration, may assist with screening and referrals by offering providers timely information and electronic referrals of patients to community organizations. Assessing patients and connecting them to relevant services is a sociotechnical process, which requires continued investment in both the social and technical components for optimal outcomes, and even when equipped with HIT, providers require training, support, and clear roles to successfully identify and respond to the social needs of patients.

Conclusion

To facilitate expansion of SDOH screening in healthcare settings, there is a need to address providers’ lack of formal training, including clarifying of roles and responsibilities required and improving handoffs within and outside of the health system.

References

A Data-Driven Algorithm to Recommend Initial Clinical Workup for Outpatient Specialty Referral

Wui Ip, MD, Priya Prahalad, MD, PhD, Jonathan P. Palma, MD, Jonathan H. Chen, MD, PhD
Stanford University School of Medicine, Palo Alto, CA

Introduction

There is a fundamental and growing gap between the supply and demand of medical expertise, as reflected in the projected shortage of 100,000 physicians by 2030.¹ The problem is particularly acute for specialty care, for which over 25 million people in the US have deficient access.² Wait times for in-person specialty visits commonly extend weeks to months after referrals are made. Adding to this problem, essential initial workup is often not completed,³ resulting in ineffective visits when the specialists do not have sufficient information to make a definitive diagnosis and treatment recommendations by the time of their first in-person visit. Such inefficiency could lead to care delay, missed opportunity to provide access to more patients, as well as dissatisfaction of patients and families.

While existing tools including electronic consults⁴ allow referring providers to solicit specialists’ opinions on the types of pre-workup needed, this remains limited in availability as it still requires a human consultant to review and respond to each request.

Our goal is to develop a different paradigm for specialty consultations with a tier of automated guides that proactively enable initial workup that would otherwise be delayed awaiting an in-person visit. Taking advantage of electronic health records (EHR) that contain thousands of specialist referral visits, we propose a data-driven algorithm inspired by Amazon’s “customers who bought A also bought B” to anticipate initial clinical workup based on how specialists cared for similar patients in the past.

Methods

Using pediatric endocrinology as an example, we extracted de-identified EHR data from a cohort of 3424 patents with new outpatient endocrinology referral at Stanford Children’s Health including > 1,150,000 instances of 8263 distinct clinical items (diagnosis codes, lab results, procedures and medications). Data were split into a training set (referral made from 2015 to 2019) and a test set (referral made in 2020). Using the training set, we calculated co-occurrence statistics of pairs of clinical items to build an item association matrix (Figure 1).

![Diagram](Image)

**Figure 1: Recommendation algorithm**

In the test set, using query clinical items (A₁, … ,Aₖ) at the time of specialty referral, the algorithm retrieves scores that resemble post-test probability from the co-occurrence matrix for all possible target items (laboratory and imaging workup orders). For each query item (A), target items (B) are ranked by estimated post-test probability P(B|A), defined as the number of patients who have query item A followed by target item B (Nₐb) divided by the number of patients with query item A (Nₐ). If a patient has q query items, q separate ranked-lists are generated. To aggregate these results, we estimate total pseudo-counts using the following equation that sums across every i-th query item: \[ \frac{1}{\sum_{i=1}^{q} N_{A_i}} \sum_{i=1}^{q} N_{A_i B} \]

Comparing to the workup items actually ordered by endocrinologists, we calculated the mean AUROC for the patients in the test set as well as the precision and recall for top 4 recommendations for these patients (4 is the average number of orders by endocrinologists). In addition, we surveyed practicing endocrinologists (n = 12) to assess the clinical appropriateness of the predicted workup orders in three common referral conditions, where we input a single query item (referral condition) to the algorithm. We also reviewed literature as external validation for these predicted orders in each referral condition.

**Results**

The algorithm achieved high discriminatory accuracy AUC > 0.9 (95% CI 0.95 - 0.96). Compared to a reference benchmark of using the overall most common orders, precision improved from 37% to 48% (p < 10⁻¹⁸) and recall improved from 27% to 39% (p < 10⁻¹⁸) for top 4 workup order recommendations when using the item-
associated recommender algorithm. In three common referral conditions (abnormal thyroid labs with high thyrotropin, obesity, and amenorrhea), the majority of the specialists considered the top recommended workup orders by the algorithm clinically appropriate (Table 1a, 1b, 1c).

### Table 1a (orange): Top recommended orders for abnormal thyroid study with high thyrotropin (n = 219)

<table>
<thead>
<tr>
<th>Orders</th>
<th>PPV (%)</th>
<th>Relative Ratio</th>
<th>Endocrine Prevalence (%)</th>
<th>Outpatient Prevalence (%)</th>
<th>% Endocrinologist consider appropriate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Thyroid stimulating hormone</td>
<td>60.9</td>
<td>1.0</td>
<td>61.7</td>
<td>17.1</td>
<td>92</td>
</tr>
<tr>
<td>2. Free thyroxine</td>
<td>60.3</td>
<td>1.2</td>
<td>56.8</td>
<td>7.5</td>
<td>100</td>
</tr>
<tr>
<td>3. Thyroglobulin antibody</td>
<td>41.7</td>
<td>3.7</td>
<td>12.8</td>
<td>0.5</td>
<td>92</td>
</tr>
<tr>
<td>4. Thyroperoxidase antibody</td>
<td>39.1</td>
<td>3.2</td>
<td>13.7</td>
<td>1.0</td>
<td>92</td>
</tr>
<tr>
<td>5. Vitamin D level</td>
<td>9.3</td>
<td>0.3</td>
<td>31.4</td>
<td>8.1</td>
<td>0</td>
</tr>
<tr>
<td>6. Serum cortisol</td>
<td>8.6</td>
<td>0.7</td>
<td>11.3</td>
<td>0.7</td>
<td>0</td>
</tr>
</tbody>
</table>

**PPV** - positive predictive value

**Relative ratio** - the ratio of the probability of the order given the query item to the probability of the order given the lack of the query item

**Endocrine prevalence** - the percentage of patients have the orders in the endocrine referral cohort

**Outpatient prevalence** = the percentage of patients have the orders among all outpatients

Table 1b (grey): Top recommended orders for obesity (n = 291)

<table>
<thead>
<tr>
<th>Orders</th>
<th>PPV (%)</th>
<th>Relative Ratio</th>
<th>Endocrine Prevalence (%)</th>
<th>Outpatient Prevalence (%)</th>
<th>% Endocrinologist consider appropriate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Hemoglobin A1c</td>
<td>40.2</td>
<td>1.9</td>
<td>22.5</td>
<td>12.5</td>
<td>100</td>
</tr>
<tr>
<td>2. Thyroid stimulating hormone (TSH)</td>
<td>28.0</td>
<td>0.4</td>
<td>61.7</td>
<td>17.1</td>
<td>75</td>
</tr>
<tr>
<td>3. Free thyroxine</td>
<td>25.6</td>
<td>0.4</td>
<td>56.8</td>
<td>7.5</td>
<td>42</td>
</tr>
<tr>
<td>4. Comprehensive metabolic panel</td>
<td>25.6</td>
<td>0.5</td>
<td>53.8</td>
<td>23.1</td>
<td>42</td>
</tr>
<tr>
<td>5. Lipid panel</td>
<td>25.0</td>
<td>1.4</td>
<td>18.3</td>
<td>12.9</td>
<td>100</td>
</tr>
<tr>
<td>6. Vitamin D level</td>
<td>20.7</td>
<td>0.6</td>
<td>31.4</td>
<td>8.1</td>
<td>42</td>
</tr>
</tbody>
</table>

Table 1c (blue): Top recommended orders for amenorrhea (n = 118)

<table>
<thead>
<tr>
<th>Orders</th>
<th>PPV (%)</th>
<th>Relative Ratio</th>
<th>Endocrine Prevalence (%)</th>
<th>Outpatient Prevalence (%)</th>
<th>% Endocrinologist consider appropriate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Prolactin</td>
<td>41.4</td>
<td>4.8</td>
<td>8.9</td>
<td>0.4</td>
<td>92</td>
</tr>
<tr>
<td>2. Luteinizing hormone</td>
<td>37.9</td>
<td>2.2</td>
<td>17.5</td>
<td>0.9</td>
<td>100</td>
</tr>
<tr>
<td>3. Follicle stimulating hormone</td>
<td>34.5</td>
<td>2.1</td>
<td>16.5</td>
<td>1.7</td>
<td>100</td>
</tr>
<tr>
<td>4. Estradiol</td>
<td>27.6</td>
<td>4.7</td>
<td>6.1</td>
<td>0.4</td>
<td>100</td>
</tr>
<tr>
<td>5. 17 Hydroxy-progesterone</td>
<td>24.1</td>
<td>2.5</td>
<td>9.7</td>
<td>0.2</td>
<td>100</td>
</tr>
<tr>
<td>6. Dehydroepiandrosterone sulfate</td>
<td>17.2</td>
<td>2.2</td>
<td>8.0</td>
<td>0.4</td>
<td>92</td>
</tr>
</tbody>
</table>

**Discussion**

Using pediatric endocrinology as an example, we developed and evaluated a recommender algorithm to anticipate initial workup needs at the time of specialty referral. Such a data-driven approach could boost the capacity of the health system by at least ensuring first specialty clinic visits are more effective while sparing specialists’ time to communicate initial workup needs. While the current algorithm is not suitable for full automation given the level of precision and recall, such an algorithm could serve as a clinical decision support tool by displaying relevant clinical orders for referring providers to make the referral process more effective. Advantages of an algorithmic decision support tool compared to building consensus guidelines among specialists include (1) scalability to answer unlimited queries on-demand, (2) maintainability through automated statistical learning, (3) adaptability to respond to evolving clinical practices, and (4) personalizability of individual suggestions.

**Conclusion**

An item-association based recommender algorithm can predict a specialist’s workup orders with high discriminatory accuracy (AUC > 0.9). Specialists indicate the top recommended orders are clinically appropriate in three common referral reasons.

**References**

Tuberculosis treatment support tools: Iterative refinement based on a mixed-method randomized controlled pilot study and usability testing

Sarah Iribarren, PhD,1 Kyle Goodwin,1 Alex Stabile,1 Alfie Vidrio,1 Rebecca Schnall, PhD,2 George Demiris, PhD3
1University of Washington, Seattle, WA; 2Columbia University, New York, NY; 3University of Pennsylvania, Philadelphia, PA

Introduction

Tuberculosis (TB) is a leading cause of death globally, surpassing HIV/AIDS and malaria, despite most forms being preventable and curable. There are ~10 million active TB cases and ~1.5 million deaths each year.1 Effective cure of TB requires a patient to take multiple antibiotics without interruption following a strict schedule for at least 6 months. Unfortunately, many fail to complete their full course successfully which leads to reduced cure rates, prolonged infectiousness, more severe disease and death, and the emergence of multi-drug resistant TB strains. If current trends continue, multidrug-resistant TB could kill as many as 2.5 million people per year and cost the global economy up to US$16.7 trillion by 2050. Innovative interventions, such as digital adherence monitoring, have the potential to enhance patient-centric quality care and help TB programs overcome some of their biggest challenges. Our team has developed digital adherence tools to empower patients and TB teams to improve patient outcomes for this long and challenging treatment. The TB Treatment Support Tools (TB-TSTs) links a patient-centered mobile application, an engineered home-based drug metabolite urine test and interactive communication with a treatment supporter. The objectives of this research were to identify issues during pilot testing and further iteratively refine the TB-TSTs using multiple evaluation methods.

Methods

We report the iterative development and evaluation of the TB-TSTs. First, we refined the tools base on pilot study findings then conducted further usability testing with a new participant population to further refine. The mixed-methods pilot study was conducted within a public respiratory-specialized reference hospital in the Province of Buenos Aires, Argentina. Patients newly diagnosed with TB who were 18 or older, and had mobile phone access were randomized to usual care (𝑛 = 21) or the intervention (𝑛 = 21). Intervention group participants were asked to report self-administering their medication daily, if they were experiencing any potential medication side effects, and submit a photo of the urine drug metabolite test 3-5 times per week using the patient app. A nurse served as the treatment supporter to the intervention cohort and responded to questions, issues, or queried participants who missed a report to ask if assistance was needed. Primary outcomes were feasibility and acceptability; secondary outcomes explored initial efficacy. Feasibility and acceptability were assessed by app usage, issues identified, and experiences using the system reported in exit interviews and survey and during the pilot study. Findings were used to iteratively refine the app and test strip. We then conducted 3 cycles of usability testing using think-out-loud protocol of the patient facing app and assess usability scores using the mHealth Usability Questionnaire (MAUQ) and the HealthIT Usability Evaluation Scale (ITUES) surveys.2,3 Further design refinement efforts were conducted through targeted directed research groups with teams of human-centered design and engineering and multidisciplinary students.

Results

For those randomized to the intervention, interactions between participants and treatment supporter focused on promoting adherence through missed report inquiries and check-ins. Although there was a decrease in interactions over time, adherence support remained needed throughout. App related support was largely due to login issues. Over half reported side effects and requested assistance. Thirty-three themes within the main categories of what worked, issues experienced, and

Fig 1. Example of home screen refinement
recommendations were identified in the analysis of the exit interviews. Early efficacy results were that higher rates of treatment success were seen in the intervention group (n=20) compared to the control group (n=17). Although not all intervention participants reported daily, none abandoned treatment. Only participants in the intervention group had follow-up sputum tests to be classified as cured. Iterative design improvements based on findings included, for example, offline app accessibility, improved and simplified reporting flow, walkthrough app orientation and videos, confirmation of

Table 1. MAUQ and ITUES subscales

<table>
<thead>
<tr>
<th>Subscale</th>
<th>Mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAUQ</td>
<td>5.96 (0.46)</td>
</tr>
<tr>
<td>Ease of use and satisfaction</td>
<td>5.9 (0.9)</td>
</tr>
<tr>
<td>System information arrangement</td>
<td>5.65 (1.15)</td>
</tr>
<tr>
<td>Usefulness</td>
<td>6.07 (0.95)</td>
</tr>
<tr>
<td>ITUES</td>
<td>4.27 (0.35)</td>
</tr>
<tr>
<td>Quality of working life</td>
<td>4.32 (0.61)</td>
</tr>
<tr>
<td>Perceived usefulness</td>
<td>4.33 (0.61)</td>
</tr>
<tr>
<td>Perceived ease of use</td>
<td>4.41 (0.52)</td>
</tr>
<tr>
<td>Control for the user</td>
<td>3.89 (.76)</td>
</tr>
</tbody>
</table>

successfully submitted reports (Fig 1). Improvements were also made to the treatment supporter interface. During the next refinement and evaluation phase, 26 participants in all three testing cycles rated the TB-TSTs application as having high usability with a mean usability score of 5.96 out of seven on the MAUQ and 4.27 out of five on the ITUES.

Discussion

Interactive communication was considered an essential intervention component. The efforts of the treatment supporter resulted in timely and effective handling of participant needs and created a sense of partnership. The iterative approaches of participant feedback from pilot testing and further usability cycle evaluations resulted in several refinements and additions to the patient facing app (Fig 2). As a next phase, the refined TB-TSTs will be evaluated in a randomized controlled trial with planned recruitment of 400 participants at four public hospitals in Argentina.

References

3. Schnall, R., Cho, H., & Liu, J. Health information technology usability Evaluation Scale (Health-ITUES) for usability assessment of mobile health technology: Validation study. JMIR MHealth and UHealth, 2018; 6(1).

Acknowledgements

We gratefully acknowledge the patients, TB teams, students, research staff who participated in the various stages and iterations of this research project. This project was supported in part by the National Institute of Nursing Research (K23NR017210: Iribarren) and National Institute of Allergies and Infectious Diseases (R01AI147129: Iribarren, Rubinstein) of the National Institute of Health.
TeamTat: A Collaborative Text Annotation Tool for Creating Gold-Standard Corpora

Rezarta Islamaj¹, PhD, Dongseop Kwon², PhD, Sun Kim¹, PhD, Zhiyong Lu¹, PhD
¹National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA
²School of Software Convergence, Myongji University, Seoul 03674, South Korea

Introduction

Biomedical research heavily depends on expert-curated data, which typically involves human annotators perusing journal articles and marking up the written facts for knowledge extraction. Such labelled datasets are incredibly valuable both for population of knowledge repositories, and for the development of AI-powered machine-learning algorithms. Human curation is a costly and time-consuming process and, given the rapid growth of biomedical literature, it is paramount to provide annotators with intuitive tools capable of speeding up the process and maintain expert subject-matter quality.

TeamTat¹ is a freely accessible annotation system that fulfills the desiderata to support multi-user projects, offer project management features, and provide an easy-to-use interface². TeamTat was developed to help database curators, and scientists needing manually curated documents, by providing an engaging annotation interface interoperable with a variety of data formats, independent workspace for annotators, collaborative workspace to resolve annotation differences, quality assessment via inter-annotator agreement, etc. Taken together, TeamTat is an all-in-one system with a set of features that cannot be found in existing tools. For example, our previous tool, ezTag³, does not support team annotation and project management. Similarly, other tools (brat⁴, PDFanno⁵, etc.) surveyed in Neves and Seva² are limited in their support of local installation, web capabilities, PubMed/PMC integration, full-text annotation, figure display, collaborative annotation, document format, etc.

Tool Description

TeamTat combines a minimalist user-interface that increases annotator’s focus and efficiency during annotation tasks. TeamTat is a multi-role system with project managers setting up an annotation team and an annotation project. An annotation project uses as input a set of documents and returns those documents with in-line text annotations, downloadable at any time. The project manager sets up the annotation schema supporting both entities, and relations, and organizes the annotation in rounds, reflecting the iterative nature of production life cycle. An annotation round allows annotators to work independently or collaboratively, and inter-annotator agreement statistics allow the evaluation of corpus quality, as shown in Figure 1.
TeamTat annotation interface is designed for ease of integration with PubMed/PMC, ease of use with full text articles and supports figure display for the annotators’ convenience. Annotators can work on the same document independently and anonymously in their workspaces, or collaboratively, with live discussions to resolve disagreements. The important features of the annotation interface in TeamTat include: (1) collaboration between annotators, (2) documents of any length, including full text journal articles, (3) optimized annotation interface for user-friendly browsing, (4) documents can be annotated for triage, for entities and for relations, (5) review interface to facilitate agreement/disagreement between annotation partners, etc.

TeamTat project management interface allows: (1) annotation projects to be organized in multiple rounds, (2) project managers to track task development and completion, (3) the assessment of annotation quality via inter-annotator agreement statistics, and (4) the creation of corpus report statistics. Project managers can define annotation and normalization options for both entities and relations, can select and distribute documents anonymously to prevent bias. Multiple users can work on a document independently and TeamTat offers visual clues to notify the manager for task completion.

TeamTat project administrators can set up TeamTat locally to accommodate data privacy concerns. Documents can be in BioC, plain text or PDF format, and Unicode support allows for documents in different languages. TeamTat interface can display all figures of PubMed Central full text articles, and it is integrated with PubMed/PMC BioC APIs\(^6,7\) to easily retrieve documents or uploaded from local repositories. Since images frequently contain crucial information or experimental evidence, TeamTat is equipped to display images for annotators’ convenience.

Usage
TeamTat offers support for annotation efficiency, consistency, scale; provides an intuitive interface; and mimics a project development workflow with clear procedures that allow the development of a gold standard corpus. TeamTat does not collect any personal information data from its users. TeamTat has been used to build several gold standard annotation corpora such as NLM-Chem\(^8\), NLM-Gene\(^9\), etc.

Availability
TeamTat is available at [https://www.teamtat.org] and in GitHub [https://github.com/ncbi-nlp/TeamTat].

Acknowledgement
This research is supported by the NIH Intramural Research Program, National Library of Medicine.

References
NLM-Chem, a Gold-Standard Corpus for Chemical Entity Recognition in PubMed Full Text Literature

Rezarta Islamaj, PhD, Robert Leaman, PhD, David Cissel, PhD, Cathleen Coss, PhD, Carol Fisher, Rob Guzman, Preeti G. Kochar, PhD, Stella Koppel, PhD, Dorothy Trinh, PhD, Keiko Sekiya, PhD, Janice Ward, Deborah Whitman, and Zhiyong, Lu, PhD
National Library of Medicine, National Institutes of Health, Bethesda, MD, 20894, USA

Introduction

Chemical and drug names are one of the entity types most frequently searched in PubMed\(^1\). Chemicals appear throughout the biomedical research literature and encompass studies beyond chemistry, such as medicine, biology, and pharmacology. Identifying chemical names is challenging because they have numerous typographical variants and synonyms, and authors frequently do not follow standard nomenclatures. Moreover, chemical names are often ambiguous, especially when abbreviated, and could refer to known compounds, a mixture of compounds, or novel, previously unknown compounds. These difficulties are compounded in articles’ full text, causing substantial performance reduction in automated chemical named entity recognition (NER) systems trained using only data annotated for titles and abstracts. The creation and maintenance of chemical terminologies and chemical corpora are significant concerns in the biomedical community. Recent efforts have developed several valuable chemical\(^2,3\) corpora for text mining, which have served as valuable resources for the text mining community in recent years. However, a corpus of annotated full-text articles focusing on chemicals does not currently exist.

Methods

NLM-Chem Corpus Development and Characteristics

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Characteristics of the NLM-Chem corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of full text articles</td>
<td>150</td>
</tr>
<tr>
<td>Number of journals</td>
<td>61</td>
</tr>
<tr>
<td>Number of total chemical annotations</td>
<td>38,342</td>
</tr>
<tr>
<td>Number of unique chemical strings</td>
<td>4,867</td>
</tr>
<tr>
<td>Number of unique MeSH assignments</td>
<td>2,064</td>
</tr>
<tr>
<td>Average number of chemical annotations per article (max)</td>
<td>256 (682)</td>
</tr>
<tr>
<td>Average number of unique MeSH assignments per article (max)</td>
<td>34 (91)</td>
</tr>
</tbody>
</table>

The NLM Chemical (NLM-Chem) corpus\(^4\) is currently the largest corpus of PubMed full-text articles annotated with chemical entities at a high degree of granularity (Table 1). The NLM-Chem corpus brings a significant enrichment compared to previously annotated corpora (a 135% increase in MeSH identifiers and a 170% increase in chemical mentions). Full-text annotation provides significantly greater language diversity than was present in titles and abstracts alone.

The development of the NLM-Chem corpus followed a systematic approach. We selected articles where manual curation would be useful for tool improvement and automated tools do not produce accurate results. Ten NLM indexers, with an average work experience of over 20 years in biomedical data curation, annotated 150 full-text articles, doubly annotating each article. The annotation load was evenly distributed, and indexers were randomly paired per article to control for annotation bias. There were three rounds of annotation: 1) Annotators working independently using pre-annotations, 2) Independent review comparing annotations between anonymous partners, 3) Collaborative discussion between partners and revision to produce 100% agreement. Inter-annotator agreement was measured after round 1 and was determined to be approximately 80%. We developed a customized annotation tool, TeamTat\(^5\), to help the indexers in this endeavor.
Results
Improved chemical automatic identification performance

Table 2 summarizes four evaluations of automated chemical recognition tools. The first evaluation shows the performance of our previous state-of-the-art tool, TaggerOne\(^6\), when trained using the existing resource (the BC5CDR corpus\(^3\)) and evaluated on the NLM-Chem test dataset. This model was trained on data from abstracts only; the data shows it is not able to recognize the greater diversity of chemical mentions found in the full text. The second evaluation shows TaggerOne trained on the BC5CDR corpus and the NLM-Chem train dataset, evaluated on the NLM-Chem test dataset. This evaluation clearly shows a sharp increase in performance obtained when trained using a combination of PubMed abstracts and full-text documents. The next two evaluations show the results of the new chemical tagger we developed based on a deep learning model and an approach we termed Multiple Terminology Candidate Resolution. The two evaluations contrast between using a title and abstract corpus versus the full-text annotations, so we can separate the improvement achieved due to the ability of the deep learning architecture to better generalize from context and the availability of full-text annotations. We can clearly see the large improvement in the precision and recall of chemical mentions, and additionally an increase in the accuracy of correctly identifying the MeSH identifiers. Note that these results are not far from the human inter-annotator agreement values.

Table 2 Text mining evaluation, showing performance results on the NLM-Chem test set, when the training is performed on the listed combinations of training sets, and controlled on the development set.

<table>
<thead>
<tr>
<th>System</th>
<th>Training Corpus</th>
<th>Chemical Mention Recognition</th>
<th>MeSH ID Recognition</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td>TaggerOne</td>
<td>BC5CDR</td>
<td>0.580</td>
<td>0.414</td>
</tr>
<tr>
<td></td>
<td>BC5CDR + NLM-CHEM</td>
<td>0.724</td>
<td>0.615</td>
</tr>
<tr>
<td>Bluebert+MTCR</td>
<td>BC5CDR</td>
<td>0.731</td>
<td>0.523</td>
</tr>
<tr>
<td></td>
<td>BC5CDR + NLM-CHEM</td>
<td><strong>0.810</strong></td>
<td><strong>0.711</strong></td>
</tr>
</tbody>
</table>

Conclusion
The NLM-Chem corpus contains 150 full-text journal articles selected both to be rich in chemical mentions and for articles where human annotation was expected to be most valuable. We built a new end-to-end chemical NER system and demonstrated improved performance in full-text articles. Finally, our new resource is publicly available at [https://www.ncbi.nlm.nih.gov/research/bionlp/Data](https://www.ncbi.nlm.nih.gov/research/bionlp/Data) and the chemical entity recognition results are available via API [https://www.ncbi.nlm.nih.gov/research/bionlp/APIs/](https://www.ncbi.nlm.nih.gov/research/bionlp/APIs/) and/or Pubtator: [https://www.ncbi.nlm.nih.gov/research/pubtator/](https://www.ncbi.nlm.nih.gov/research/pubtator/).

References
Medical Curriculum Elective to Integrate Clinical Analytics and Artificial Intelligence in Preparing Physicians for the Digitally Enabled Practice

Miriam Isola, DrPH\textsuperscript{1}, Radhika Sreedhar, MD\textsuperscript{1}, Linda F. Chang, PharmD\textsuperscript{1}, Tushar Patel, MD\textsuperscript{1}, Max C. Anderson, PhD\textsuperscript{1}, Jacob Krive, PhD\textsuperscript{1,2,3}.

\textsuperscript{1}University of Illinois at Chicago, Chicago, IL; \textsuperscript{2}NorthShore University Health System, Evanston, IL; \textsuperscript{3}University of Chicago, Chicago, IL.

Introduction

Despite setbacks and challenges, recent advances in healthcare digital technologies and artificial intelligence (AI) demonstrated that AI in the assistive role has its place in medical practice\textsuperscript{1,2}. Consequently, there is a need to prepare medical professionals to employ AI and non-AI analytics in patient care\textsuperscript{2}. Additionally, the Accreditation Council for Graduate Medical Education (ACGME) places strong emphasis on care quality improvement (QI) and patient safety, both of which may benefit from clinical analytics and augmented intelligence to identify issues. More sophisticated AI of the future may lead to new cures for various medical conditions, thus further cementing the role of analytics and AI in medical practice. Physicians must understand strengths and limitations of predictive analytics and AI, and when and how it is appropriate to apply these technologies in patient care as well as organizationally in the acute patient care settings\textsuperscript{3,4}.

The current medical school curriculum does not prepare students for these new realities of digital medical practice and does not enable physicians to become active participants in the technology adoption process. We share results of implementing an innovative modular virtual elective in the College of Medicine that integrates Clinical Analytics and AI instruction with the existing medical school curricula, helping address this medical education gap.

Methods

This AI elective is offered to University of Illinois medical students in their final year of study. It is unique in design, as it requires no prior mathematical, computer science (CS), or statistics background. Rather than focusing on the underlying concepts behind complex AI applications, this course emphasizes physician roles in leading and adopting changes brought about by these AI algorithms. The virtually delivered course consists of 4 one-week blocks integrating clinical analytics, decision support systems, and AI with real-life practice scenarios from: (1) pathology, (2) QI, (3) Remote Monitoring and mobile health and (4) real world evidence (RWE) and patient safety. Each scenario includes lectures/discussion of critical analytics and AI concepts such as - how to develop a QI measure, algorithms and how they work, value-based care, integration of data from pharma and hospital environments. Students are divided into groups that remain static for the duration of the course, and every week has a similar structure to help students remain organized while facing topics unfamiliar to them.

Each weekly unit consists of: (1) recorded lecture and announcement outlining goals, learning objectives, and assignments; (2) group project focused on one of the four above-listed areas of AI; (3) multiple choice assessment quiz testing theoretical knowledge acquired from the textbook and other materials utilized in the course; (4) weekly reflections polling students in regards to application of their learning and assisting them with making connections between the week’s materials/assignments and professional practice. Students are challenged with hands-on assignments emphasizing physician role in providing clinical expertise for the AI and clinical analytics solutions as well as practicing professional leadership in each domain area. These aims enable students to practice everyday application of technology in medical practice and explore potential for assuming specific technology responsibilities in the healthcare organizations.

The interprofessional collaboration includes faculty from Medicine, Pathology, Health Informatics, Pharmacy, and Instructional Design each contributing in their area of expertise. Students develop competencies from online lectures and assigned readings and detailed feedback from faculty teams. Students work in groups to complete challenging assignments addressing real problems and solving some of the most relevant dilemmas in medicine today. By completing the assignments and consuming the required materials, the students are able to:
Interpret and explain foundational concepts of clinical analytics, decision support, and AI
Analyze the linkages between evidence-based medicine (EBM), medication safety, and predictive analytics
Employ AI to challenge traditional concepts of home health care by introducing cutting edge remote monitoring systems and data analysis algorithms
Develop their own clinical solution with a core AI component to augment population health
Apply algorithms to design clinical analytics systems for patient safety
Design data integration methods to produce actionable information and new clinical insights
Build real world scenarios of applied AI in Value Based Care.

Faculty are also divided into teams and collaborate weekly to review student assignments. Given the uniqueness of each project and the degree of flexibility allowed to construct innovative solutions by students, faculty have interesting opportunities to learn and debate student work.

Results
25 students from three different University of Illinois campuses completed the elective to date. Students demonstrated theoretical knowledge improvement with an average quiz score of 97% and achieved an average project assessment score of 89% reflecting development of hands-on skills. Student reflections expressed confidence in their ability to articulate ideas of implementation of AI in patient care settings, demonstrated appreciation for the role of AI in medicine, and gained insights into the intersection between the process of “quantification of medicine” and QI/patient safety.

Discussion
One of the biggest obstacles to AI instruction in medical schools is full integration of content in the daily medical practice. We present a unique modular and accessible course that can be adapted and embedded into any medical curriculum. This course demonstrates that AI can be taught to students without math and CS backgrounds, by choosing relevant topics of value to clinicians and engaging them exactly where real-life scenarios call for physician involvement with digital health technologies. The improvement in knowledge and skills we observed was impressive for a short course. Student reflections demonstrated maturity of thought in how they can apply the latest advances in AI and clinical analytics to their daily medical practice routines.

Lessons learned from the pilot course offerings helped faculty retune the curriculum to address challenges medical students face due to lack of technology coverage in their standard curriculum. Some technology areas assumed to be self-explanatory require more support, while students also expressed interest in some extra-credit opportunities related to value-based care and understanding of the financial landscape of healthcare in light of analytics.

References
Leveraging precisionFDA and Synthetic Data to Improve Veteran Healthcare

Bocheng Jing, MS1*, Anish Prasanna BS2*, Elaine Johanson, BS3, Holly Stephens, PhD2, Samir Lababidi, PhD3, Ezekiel Maier, PhD2, Amanda Purnell, PhD4
1University of California, San Francisco, Division of Geriatrics, San Francisco, CA 94115
2Booz Allen Hamilton, Bethesda, MD, 20814, USA
3U.S. Food and Drug Administration, Silver Spring, MD, 20903, USA
4Veterans Health Administration Innovation Ecosystem

*Authors contributed equally to this work

Abstract

In response to the novel coronavirus disease 2019 (COVID-19), the Food and Drug Administration (FDA) and the Veterans Health Administration (VHA) launched the COVID-19 Risk Factor Modeling Challenge to better understand the disease’s impact on the Veteran community. This challenge assessed the usefulness and improved our understanding of synthetic data for Machine Learning development of models addressing a real-world problem.

Introduction

The Veteran population has a higher prevalence of underlying risk factors for severe COVID-19 illness, such as obesity, heart disease, and diabetes1. While a better understanding of these underlying factors is gained by building predictive models with Veterans health records, the data itself is not easily accessible for research purposes. Therefore, in order to better understand risk and protective factors in this population, synthetic data was utilized for a Veterans Health Administration (VHA) Innovation Ecosystem and Food and Drug Administration (FDA) COVID-19 Risk Factor Modeling Challenge. This challenge, hosted on the secure collaborative cloud-based precisionFDA platform, called on the community to leverage Machine Learning (ML) to develop and evaluate computational models to predict COVID-19 related health outcomes in Veterans in Phase 1. In Phase 2, the challenge team assessed whether synthetic data is a reliable resource for researchers outside the VA to use for model prediction.

Methods

The first challenge phase asked participants to develop models for predicting COVID-19 outcomes, including COVID-19 status, alive or deceased status, ventilation status, number of days hospitalized and number of days in the Intensive Care Unit (ICU). Synthetic health records for 147,451 synthetic patients, generated using the Synthea synthetic patient generator2, was provided to participants for modelling. This synthetic data included conditions, encounters, observations, medications, procedures, and patient demographics. The synthetic data was split, with 80% provided to participants for model training, and 20% withheld as a test set for model evaluation. Model performance was evaluated separately for each of the five target COVID-19 outcomes. COVID-19 status, alive or deceased status, and ventilation status predictions were evaluated using the Area Under the Receiver Operating Characteristic (AUROC) metric. Predictions of days hospitalized and days in the ICU were evaluated using Root-Mean-Square Error (RMSE) and concordance index (c-index).

During Phase 2, top performing participants retrained and evaluated their models on two additional health record datasets, a second synthetic dataset generated by a Generative Adversarial Network (GAN), and de-identified real Veteran health records. Model performance on all three datasets was compared through test AUROC for binary health outcomes, and test RMSE and c-index for continuous health outcomes.

Results

The first challenge phase received 34 submissions, and a wide array of ML techniques were utilized, including Random Forest, Adaptive Boost (AdaB) Neural Network, and Ensemble approaches. Median model test AUROCs across all submissions outperformed a random prediction submission for the alive status and ventilation status health outcomes by .192 (.740>.548) and .145 (.668>.523) units respectively. Each of the top 3 submitters incorporated Gradient Boosting models for health outcome predictions. Across all submissions, model performance was generally better in predicting severe health outcomes (e.g., ventilation status, days in the ICU).
The top three participants in Phase 1 challenge were selected for Phase 2. These submissions, when trained and tested on GAN generated data, scored significantly higher in predicting COVID-19 status, and had similar results as compared with Synthea generated data across all other outcomes. Like phase 1, submissions markedly surpassed metrics generated by random guessing (.548 alive status, and .668 for ventilation status) Additionally, model performance comparatively outperformed in predicting severe outcomes such as Death Status and Ventilation Status.

Table 1. GAN Phase 2 Test Metrics for Top Performers

<table>
<thead>
<tr>
<th>COVID-19 Health Outcome</th>
<th>Median (Top Performer)</th>
<th>Synthea Synthetic Data</th>
<th>GAN Synthetic Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>COVID-19 Status</td>
<td>.517</td>
<td>.700</td>
<td></td>
</tr>
<tr>
<td>Ventilator Status</td>
<td>.778</td>
<td>.776</td>
<td></td>
</tr>
<tr>
<td>Death Status</td>
<td>.831</td>
<td>.811</td>
<td></td>
</tr>
<tr>
<td>Days in Hospitalization (RMSE)</td>
<td>6.008</td>
<td>6.583</td>
<td></td>
</tr>
<tr>
<td>Days in ICU (RMSE)</td>
<td>1.602</td>
<td>1.610</td>
<td></td>
</tr>
</tbody>
</table>

We are currently testing the top performing models on real Veteran’s health data and results are forthcoming.

Conclusion

This cross-agency collaborative challenge assessed the usefulness of synthetic data for developing AI/ML models for addressing a real-world problem. In Phase 1, participants developed models trained on synthetic health records to predict COVID-19 outcomes. Phase 1 results stimulated discussions around the efficacy of synthetic data, which Phase 2 seeks to explore by comparing models trained on real Veteran health records to models trained on synthetic GAN generated and Synthea generated health record data. Models trained on synthetic GAN health records and Synthea health records have demonstrated similar results across all outcomes, except for COVID-19 status, which models trained on GAN synthetic data score higher on, suggesting that synthetic data generation methodologies can create similar COVID-19 health outcome model performances. Top performer models are currently being tested on real Veteran health data, and results will elucidate the relationship between the efficacy of synthetic data in predicting COVID-19 health outcomes in the Veteran population.

References

Responding to a Crisis of Veteran Suicide QUICKly: A Qualitative Interdisciplinary Collaboration

Andrea F. Kalvesmaki, PhD1,2, Alec Chapman, MS1,2, Kelly S. Peterson, MS1,2,3, Mary Jo Pugh, PhD1,2, Makoto Jones, MD1,2, Theresa C. Gleason, PhD4

1Salt Lake Veteran Affairs Medical Center (SLVAMC); 2University of Utah Division of Epidemiology; 3 Veterans Health Administration Office of Analytics and Performance Integration; 4Department of Veteran Affairs, Clinical Science Research & Development Service (CSRD)

Introduction: U.S. adult suicide rates have almost doubled from 2005-2017, with rates of Veteran suicide approximately 1.5 times higher than the general population. In early 2019, an Executive Order (EO) “PREVENTS” established an interagency task force to outline a national strategy to reduce Veteran suicide. An online survey was developed to collect data on Veteran suicide intervention perceptions by the public, stakeholders and organizations working with Veterans. The survey, comprised of 21 open-ended questions, was administered May-August 2019. The task force asked a team of interdisciplinary researchers affiliated with the Veteran Affairs (VA) to analyze the survey responses. Due to the pace of decision-making to enact the EO, the analysis had to be complete within months. This led to the development of a process for computer-assisted qualitative analysis: Qualitative Interdisciplinary Collaboration (QUICK). QUICK combined qualitative analysis and natural language processing (NLP) to accelerate the analysis of a large corpus of survey data in a short amount of time, in order to inform policy efforts. This methodology could be applied to other large-scale qualitative research efforts.

Methods: Our team developed a technology-assisted qualitative analytic method combining NLP text mining techniques guided by in-depth qualitative review, with experts in Veteran health, military culture, and risk factors for suicide. Qualitative analytic methods such as sensemaking2, grounded theory3, and sensitizing concepts3,4 grounded in qualitative health research theoretical frameworks of narrative inquiry,5 the medical explanatory model6 (to uncover less common topics), as well as the United Medical Language System (UMLS)7 (to identify more unified concepts) informed the text mining process. The corpus of survey data was analyzed using a three-stage approach to 1) explore the data, 2) define and extract concepts, and 3) provide a high-level synthesis and summarization of the findings. In Stage 1, the NLP analyst used medspaCy, a library built in the spaCy framework for clinical NLP, to identify term counts and phrases8, and Latent Dirichlet Allocation (LDA)9 to produce topic models for expert review. Data was jointly coded by the qualitative and NLP analysts using the annotation tool eHost12 to review word snippets and terms and to refine topics. In Stage 2, Python packages10, 11 were used to generate interactive visual representations of the topics to review themes and topics to ensure saturation of the data. Unlike typical annotation, this joint coding process allowed the research team to rapidly analyze the corpus of data to identify key terms and topics as the theoretical grounding was applied. In Stage 3, the qualitative analyst summarized and organized text by theme and theoretical base to either confirm relatedness to known research in Veteran suicide prevention or highlight new information.

Results: A total of 9040 free-text entries from 114 organizations and 608 individuals were analyzed in three months using this approach. Response findings mapped onto two major theories in suicide prevention research. Descriptions of risk factors associated with suicide were consistent with fluid vulnerability theory, which posits that a person’s potential risk for suicide at any given point in time changes depending on circumstance 13. Findings related to descriptions of prevention and interventions echoed the approach used in the crisis risk planning model (CRP)14. Other themes such as “barriers to care,” and “data collection and collaboration” were identified within the responses. Systemic differences between organizational and individual responses indicated that many organizations were more likely to describe specific interests in therapy modality, and a lack of funding and resources being barriers to effective treatment. Individual respondents were more likely to describe a story that began at the time of separation from the military. They were also more likely to describe the complicated nature of access to care and a lack of services tailored to Veteran and Military culture.
Discussion: The need for a rapid analysis of a large body of qualitative data collected to respond to a national crisis led to a novel process of applying NLP to assist a large-scale qualitative analysis grounded in traditional qualitative methods. The data findings provided the national suicide task force insights they were able to use to meet the mandate to inform a national strategy. On June 17, 2020, the President’s Roadmap to Empower Veterans and End a National Tragedy of Suicide was published outlining policy and practice strategies based in part on the results of this analysis. This analytic process could provide a useful technique for large-scale and timely qualitative analysis. The stages of analysis and process of joint coding by the qualitative and NLP analysts allowed our team to analyze a large corpus of text with the sensitivity of traditional qualitative methods while doing so at accelerated pace. The QUICk method outlined here was submitted for VA invention disclosure (VA ID 2020-530) in 2020, and a publication is being developed. We present this process to accelerate knowledge across diverse disciplines and fields where combining qualitative and computational methods could accelerate analytic methods and produce results grounded in data derived from personal experience. From a very practical perspective, it offers a way forward that can be applied broadly to open-ended surveys such as described here, and which has application to tasks soliciting feedback from a diverse audience of stakeholders.

Acknowledgement: This material is based upon work supported by a Clinical Science award, 829-AA-null-38970, from the Department of Veteran Affairs, Veteran Health Administration, Office of Clinical Science Research and Development and is the result of work supported with resources and the use of facilities at the Veteran Affairs Informatics, Decision-Enhancement and Analytic Sciences (IDEAS) Center of Innovation, Salt Lake City, Utah.

Disclaimer: The contents do not represent the views of the U.S. Department of Veteran Affairs or the United States Government.

References

13 Rudd, M. D. (2006). Fluid vulnerability theory: A cognitive approach to understanding the process of acute and chronic suicide risk
Improving Outcomes for Pediatric Patients with Metabolic Conditions in the ED using User Centered Design for Order Sets

Swaminathan Kandaswamy, PhD¹, Shabnam Jain, MD¹,², Dwight Chambers, MD³, William R. Wilcox, MD, PhD¹,², Beesan S. Agha, DO²,⁴, Sara P. Brown, RN², Evan W. Orenstein, MD¹,²

¹Emory University, Atlanta, GA; ²Children’s Healthcare of Atlanta, Atlanta, GA; ³Boston Children’s Hospital, Boston, MA; ⁴Pediatric Emergency Medicine Associates, Atlanta, GA

Introduction

Patients with metabolic diseases are at risk of rapid decompensation due to acidosis, hyperammonemia, or hypoglycemia. Timely interventions such as dextrose containing fluids can reduce deterioration and prevent ICU admission.¹ Early identification and rapid management of these patients can be challenging, especially during high volume periods in the ED and lead to delays in treatment initiation. To improve the care process, we used User Centered design (UCD) to implement Clinical Decision Support by creating: (1) an alert to notify providers about the patient’s high-risk status, and (2) an Order Set to place appropriate orders more easily. The aims of the CDS intervention were to (1) Reduce ICU admissions from ED and (2) Decrease median time to IV fluids.

Methods

 Phase 1: We first developed an order set with the help of specialists familiar with various metabolic disease conditions. Next, we adopted UCD and performed formative usability testing. End-users were given clinical scenarios for a patient with a metabolic disease diagnosis with 3 different presentations based on acuity and need for intervention. Case A: Patient with complaint of ear pain and diagnosis of acute otitis media who does not require IV fluids, Case B: Patient with complaint of vomiting thought to be due to an infectious GI problem requiring IV fluids and labs, and Case C: A septic listless patient who requires immediate IV fluids. Users of the order set were asked to place orders in a simulated test environment and asked to think aloud as they completed the task. We also obtained qualitative feedback at the end of the session. We iteratively improved the design of the Order Set based on participant feedback until no new input was identified with 2 consecutive participants.

 Phase 2: After implementation of the Order Sets, we compared the outcome measures (percent of patients admitted to the ICU) and process measure (median time to dextrose containing fluids).

Results

 Phase 1: A total of 6 providers (4 Attendings, 2 Residents) participated in the formative testing. We made 6 design changes based on formative testing. The design changes and rationality for the changes are shown in Table 1. The alert is shown in Figure 1. The final order set design is shown in Figure 2. Participants gave positive feedback about the alert language; appreciated the repetition of the alert criteria on the order set, and master list of relevant diagnoses.

Table 1. Order Set issues and design changes

<table>
<thead>
<tr>
<th>Issue Identified</th>
<th>Design Change</th>
<th>Rationale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participants accepted order set recommendations and ordered fluid in Case A when these were not required.</td>
<td>Made orders into a panel that encouraged providers to a priori make decision on need for IV fluids based on patient clinical appearance and hydration status (see figure 2)</td>
<td>Enable decision making based on what users know – providers are more likely to know patient’s clinical appearance and oral intake status than knowing if fluids are required for specific conditions.</td>
</tr>
<tr>
<td>Participants placed fluids as 1x Maintenance Rate instead of 1.5x Maintenance Rate (MR).</td>
<td>Provided rate reference: fluid specific instructions asking them to order D10 fluids at 1.5x MR.</td>
<td>Reduce reliance on provider memory.</td>
</tr>
<tr>
<td>Participants voiced uncertainty about disease specific workups and labs required for patients.</td>
<td>Included a list of all relevant labs specific to the disease after consultation with experts.</td>
<td>Reduce reliance on provider memory. Allow recognition instead of recall.</td>
</tr>
</tbody>
</table>
Missing orders in the Order Set
(1) Change order to Venous iSTAT for gas. (2) Added Normal Saline Bolus (3) Changed carnitine dosing instructions to single dose
Include all relevant orders in the order set and default order details matching ED requirements.

Participants placed duplicate orders as there was a pre-selected diagnosis (order panel) and they re-selected the diagnosis in the master list of metabolic conditions.
Included only either pre-selected diagnosis or master list of diagnosis.
Design to eliminate errors.

Participants felt forced to open order set out of concern for not being able to get valuable information before they were ready. They did not notice Acknowledgement reason (“Remain active for me”) applicable to this concern.
Included acknowledgement reason “Yet to see the patient, show again later.”
Match design with users’ mental model

Figure 1. Alert for patients at risk of metabolic decompensation.
Figure 2. Order set Design.

Phase 2: There were 3055 eligible encounters prior to intervention (May 01 2009 - Sep 15 2020) of which 1525 got admitted; and 185 encounters post intervention (Sept 16 2020 to – May 31 2021) of which 109 got admitted. The alert acceptance rate was 44%. Post-intervention, ICU admission rate decreased from 32% (489/1525) to 26% (28/109) p 0.2. The median time to dextrose fluids administration decreased from 172 minutes to 104 minutes (p < 0.01).

Conclusion
UCD helped ordering providers recognize metabolic disease faster and know what to do for different metabolic diseases. Formative usability testing led to design changes relative to design based on CDS design principles alone. CDS design can be improved by (1) enabling decision making based on what providers know e.g. whether patient is ill appearing and not details of underlying disease (2) reducing reliance on provider memory and allowing recognition instead of recall e.g. providing disease specific recommended labs and (3) changing representation to match provider mental models e.g. rewording acknowledgement reasons.

References
How Robust is Your Risk Model? Assessing Subpopulation Performance Heterogeneity in Risk Models Based on Discrimination, Calibration, and Fairness Measures: An Atrial Fibrillation Use Case

Uri Kartoun PhD\textsuperscript{1}, Shaan Khurshid MD\textsuperscript{2,3}, Bum C Kwon PhD\textsuperscript{1}, Amit V Khera MD MSc\textsuperscript{2,3}, Patrick T Ellinor MD PhD\textsuperscript{2,4}, Steven A Lubitz MD MPH\textsuperscript{2,4}, Kenney Ng PhD\textsuperscript{1}

1. Center for Computational Health, IBM Research, Cambridge, MA, USA.
2. Cardiovascular Disease Initiative, Broad Institute of the Massachusetts Institute of Technology and Harvard University, Cambridge, MA, USA.
3. Division of Cardiology, Massachusetts General Hospital, Boston, MA, USA.
4. Cardiac Arrhythmia Service, Massachusetts General Hospital, Boston, MA, USA.

* Authors contributed equally.

**Introduction:** Atrial fibrillation (AF) is a cardiovascular condition linked to life-threatening outcomes, such as heart failure (HF) and stroke. Using mechanisms such as the electronic health record-AF (EHR-AF) score \cite{1} can assess a patient’s risk at the point-of-care to help physicians more accurately apply a variety of preventive strategies. Such strategies include, for example, weight-loss interventions and blood pressure control as well as inform screening to detect asymptomatic AF earlier. Widely used metrics of risk model performance (e.g., discrimination, calibration) as well as more novel ones such as fairness-based metrics to assess biases for underprivileged subpopulations are commonly applied to an entire eligible population (e.g., patients 45 years old or older). However, when applied to specific subpopulations (e.g., different age ranges, sex, races, and prevalent comorbidities), the measures may yield important differences in risk score performance. In this study we aimed to raise awareness about the robustness and applicability of risk models by characterizing performance heterogeneity across subpopulations of clinical relevance on the EHR-AF risk score.

**Methods:** From the 502,521 patients enrolled in the UK Biobank, we included the subset of patients aged 45 years or older at enrollment, without a history of AF and with available height, weight, and systolic and diastolic blood pressure measurements. These 4 measurements, along with additional covariates, such as sex, race, and comorbidities, were used to calculate all patients’ EHR-AF scores at enrollment. The patients were then followed for up to 5 years (consistently with EHR-AF’s original design) to assess associations between their EHR-AF scores and incident AF. Analyses included stratification of the population into a variety of subpopulations, including by sex, race, prevalent comorbidities including HF and/or stroke, and age strata (45–49, 50–54, etc.). Common performance measures were then applied to all combinations of subpopulations (as well as the population as a whole). These included traditional measures, such as concordance indices, calibration slopes, and standardized hazard ratios. To identify additional differences between the subpopulations within the context of possible underprivileged subpopulations (white versus nonwhite, males versus females), we further applied fairness-related measures, including the statistical parity difference and true positive rate difference \cite{2}. To calculate confidence intervals, we used bootstrapping.

**Results:** We identified 445,329 patients who met our inclusion and exclusion criteria. Notably, discrimination of incident AF for patients with prevalent HF was the lowest, secondary to stroke and the overall population (Fig. 1). With respect to calibration, HF was the only subpopulation in which AF risk was overestimated using the score. The differences across races were minor, and risk appeared underestimated among females as compared to males. Note that we conducted additional analyses, and the accompanying charts are available and include over 100 subpopulations stratified by age ranges, races, and sex, highlighting intriguing conclusions. Regarding the assessment of fairness measures (Fig. 2), notably, EHR-AF generates higher risk scores for males compared to females in the age range of 55–69. Additionally, the true positive rate was lower for females compared to males in the same age ranges. For the extreme age ranges, however, there was little to no bias.

**Discussion:** This work reveals potential robustness limitations in a well-validated clinical risk model. We observed that performance appears to vary substantially when analyzed within specific subpopulations as opposed to the overall population. Understanding these model behaviors can help clinicians know when it is appropriate to apply the risk model. Assessing these limitations can also help improve model robustness by targeting mitigation strategies such as recalibration or retraining to certain subpopulations which we plan to do in subsequent studies. Similarly, calculating fairness-related measures to identify potential biases may be more useful if applied to specific subpopulations because
such a methodology may be capable of identifying specific subpopulations with biases versus those that are free of bias.

References

Fig. 1. Commonly used performance measures applied to the overall population as well as to subpopulations. Prev. = Prevalence; Inc. = Incidence per 1,000-patient years. Concordance index values typically range between 0.5 and 1.0. An optimal calibration slope equals to 1; values above indicate underestimation and values below indicate overestimation of the risk.

Fig. 2. Fairness-related measures applicative to sex stratified by age ranges. EHR-AF = 7 (median) was selected as the classifier threshold. Inc. = Incidence per 1,000-patient years.
Informatics-Driven COVID-19 Vaccine Enterprise: Lessons for the Future

Nitu Kashyap, MD FAMIA1,2; Sarah Kelly, PharmD MHA1, Rebecca McCray, MS BSN RN1, Allen Hsiao, MD FAAP1,2
1Yale New Haven Health, New Haven, CT; 2Yale School of Medicine, New Haven, CT;

Abstract

Connecticut (CT) has sustained high rates of COVID-19 vaccinations as one of top 3 States in the country1. We share experience from a large academic medical center that used EHR and informatics tools to support a vaccine enterprise with rapid ramp up, high volumes, low waste and high safety track that was recognized by the CDC. We have administered ~440,000 doses with a waste rate of ~0.1%.

The objective of this presentation is to understand challenges in setting up a large-scale vaccine administration campaign and to learn best practices for future similar campaigns for other public health and population health issues.

The Emergency Use Authorization (EUA) for COVID-19 vaccines was a welcome development in the fight against the pandemic. The unprecedented public health response also required health systems to stretch their resources and play a key role2. In CT, the State Department of Health (DPH) managed vaccine allocations, while distribution and patient connections were managed by healthcare delivery organizations. Key elements of this vaccine enterprise included a robust pharmacy and supply chain to manage cold chain, PPE and other supplies. The CDC site visit team noted Yale’s process for cold chain monitoring as a best practice. Efficient self-service scheduling via a web-based screening decision tree that performed all safety checks prior to scheduling allowed us to manage throughput and social distancing at vaccination sites. Several Clinical Decision Support (CDS) tools to identify correct intervals with any prior infection or treatment, correct vaccine type, dose intervals, allergies, age restrictions and reactions to prior dose were developed to allow for a safe mass vaccination program. The same systems also helped in early detection and management of conditions observed post vaccination3.

We created a process of bidirectional vaccine information exchange with real time submission of all administrations and nightly queries (Figure 1) to verify the up-to-date vaccine status of individuals scheduled for vaccination. This helped anticipate no shows to manage waste and get more shots into arms. CDC noted that this was a novel approach of data driven vaccine safety. Robust real-time analytics on vaccination trends, schedule utilization, population outreach, potential no shows, vaccine mismatch were tools utilized at all levels of the enterprise from planning to operations to call center.

Results

Across the vaccine enterprise at Yale New Haven Health, 443,135 doses have been administered to 232,211 unique patients. Vaccination rates ramped up rapidly after the Pfizer vaccine EUA was granted and have rapidly tapered off since June. CT continues to maintain above average rates for vaccinating its population.

Conclusion

Rapidly creating a large vaccine enterprise to safely administer covid vaccines in an equitable manner requires a strong informatics infrastructure that supports all aspects from planning and execution to descriptive analytics and interoperability across multiple systems. This is a great example of public-private partnership highlighting the importance and use of standards.

References

NLP-Driven Automated News Monitoring Solution for COVID-19 Response

Mary Kennedy-Moore¹, Stephen A Mustillo¹, Conor Kennedy¹, Zakir Hussain¹, Jessica D. Tenenbaum, PhD²,³

¹Ernst & Young, New York, NY, ²Duke University, Durham, NC, ³NC Dept. of Health & Human Services, Raleigh, NC

Introduction
In the current information age, stories on new COVID developments are often released to the media right away, sometimes before the State has been notified of the situation.¹ This lag in information may prevent public health officials from responding proactively. Stories are often posted on local town or county news sites, rather than well-publicized national sites, and there is a considerable amount of noise in the news on COVID, making manual news searches time consuming, and neither scalable nor exhaustive. Beyond actual news content, it is often useful to note overall sentiment of coverage, whether in mainstream or social media.²⁻⁴ Therefore, the NLP-driven Automated News Monitoring Solution was developed to scan local and national news sites, and automatically surface and analyze gathered insights at scale. This approach was applied to two different use cases to support the State of North Carolina’s COVID-19 response.

Methods
The first use case was identification of COVID-19 clusters/outbreaks across the state. In late Summer 2020, with daily new cases continuing at an alarming rate, return-to-school considerations on the horizon, and a continued media focus on the pandemic, local news outlets in particular were rapidly publishing breaking news stories on new groupings of COVID-19 cases. Often, these news stories would break before the state was made aware of a cluster/outbreak via traditional monitoring processes.

The second use case was regional vaccine supply and demand alignment. Again, monitoring the news manually would be impossible to do exhaustively at scale. Instead of simply identifying certain entities from news articles, the State needed a way to quickly digest the news and understand if it was positive or negative, and anything that could help inform allocation decisions to vaccination providers across the state.

Both use cases feature a similar technical structure and process:

1. **Gathering the news data.** A wide net is cast to gather all news that could potentially be relevant to the given use case, leveraging a leading consumer intelligence tool’s API. The titles are saved and a web scraper used to pull the full text of the articles, along with other pertinent data.

2. **Preprocessing and tagging data.** The data are cleaned, and the article text matched against sets of keywords that help identify the location (e.g. city, county) and setting (e.g. school, church, etc.) of an outbreak for the Clusters/Outbreaks use case, and the topic and demographic of an article for Vaccines use case.

3. **Entity extraction and categorization (Clusters/Outbreaks use case only).** NLP methods are employed to extract entities from the articles – such as “Duke University” – which provide a stronger basis for classifying both the setting and relevance an article, as well as aiding in identifying the county in which the article takes place.

4. **Contextualized Sentiment Scoring (Vaccines use case only).** Different from typical sentiment analysis, the headlines of articles were run through two models – one trained to evaluate how well the State was handling the vaccine rollout, and another trained to evaluate how willing the populace were to get vaccinated. Both output a score between -1 (bad news) and 1 (good news).

5. **Filtering.** A set of decision trees based upon the results of the keyword tagging and setting extraction is employed to filter the set of articles. A different set of keywords and tags is used for each use case. For the Clusters/Outbreaks use case, a logistic regression model trained on a set of manually tagged data is employed to aid in the classification of these articles. This system achieves 98% accuracy, and an F1 score - more useful due to the large number of True Negatives - of 91% against manually classified data.

6. **Grouping of similar articles.** For a succinct and useful update, NLP methods are used to group articles that are reporting the same story – both in the same day and over the past week – so that each story is only reported once.

7. **Automated email and further use of data.** For the Clusters/Outbreaks case, an email that offers a succinct bullet-point list of articles reporting new clusters/outbreaks along with the setting of the story and a link to the article is automatically generated and distributed. For both cases, the daily data is also loaded into a dashboard for the respective use case for further insight exploration and use by the State.
Results
For both use cases, the Automated News Monitoring solution enables gathering and tracking of trends to inform key response activities. The dashboards provide individual links to relevant news stories – providing both details and transparency – as well as analysis of trends across categories and locations over time. (See Figure.) The automated cluster/outbreak identification allowed the state to immediately respond and allocate resources (e.g. contact tracers) where applicable, as well as to track trends of cluster/outbreak settings overall to inform prevention measures. This complements and supplements the tracking that is already done through required public health reporting of COVID-19 cases. The vaccines use case allowed the state to monitor public perception of the vaccine rollout, as well as track willingness of certain populations and regions. This helped support vaccine allocation decisions, including quick notification of distribution bottlenecks and planning of community vaccination events.

Discussion
The Automated News Monitoring Solution was developed to address the need to keep abreast of new clusters and outbreaks in near real-time as a complement to mandatory public health reporting. The solution also helped to promptly identify and track any news, emerging complaints, events, and bottlenecks related to the State’s performance in vaccine distribution, as well as regional willingness of State residents to take the vaccine, thus impacting demand. Through the application of a variety of data wrangling, data analysis, NLP, and automation techniques a generalizable and automatable framework was created that can be applied both to these same use cases in other states, and to other use cases across industries. This solution can be easily scaled to handle large volumes of articles and modified to provide targeted insights on different topics, with links to the tagged article titles and body texts themselves. It can also be easily applied by other states to the same pressing issues or generalized by other actors to fit their needs and – with minimal human input – provide a highly accurate and succinct output.

References
COVID-19 Clinical and Social Surveillance System for After School Programs

Ellen K Kerns, PhD MPH¹, Russell J. McCulloh, MD¹, Chad J. Abresch, PhD¹
¹Dept. of Pediatrics, University of Nebraska Medical Center, Omaha, NE, United States

Introduction

The COVID-19 pandemic has created medical, social, and economic hardships for families, particularly in communities of color. School based programs play a key role in addressing household challenges; however, there is a critical need for scalable tools schools can use to keep their facilities safely open and adequately respond to household challenges. We designed an mHealth program to fill this need by providing after school programs an app with two decision-support tools: a daily COVID-19 Household Symptom Checker and a bi-weekly Household Challenges Tracker. Our overall program includes: 1) app-based feedback for families, delivering current guidance from their local health department and CDC; 2) custom dashboards for the after-school programs, summarizing households’ responses to inform infection mitigation strategy; and 3) community navigators, who expedite public aid and community assistance to requesting families. Our research objective is to measure the program’s feasibility and acceptability by piloting it in two after school programs in Omaha, NE: NorthStar and Girls Inc.

Methods

A flow diagram of the mHealth program components is presented in Figure 1. Each after school program received a health department approved daily symptom screening tool (COVID-19 Household Symptom Checker) and a bi-weekly household challenges screener (Household Challenges Tracker) accessible via website or by app customized with their local context (colors, logo, distance learning/attendance procedures, etc.). Prior to release, the tools were usability tested with families from participating schools prior to release and were revised based upon feedback from these families, the after-school program administrators, and the family navigators. Both tools were provided in English, Spanish, and Nepali. Household preferred language, contact information, and student demographics are entered upon first use (Figure 2A). Tool translations were provided by certified medical translators familiar with the community served by the after-school program. After this initial registration, households are presented with the two tools upon opening the app or website (Figure 2B). The symptom checker uses health department guidelines to advise staying home or go to school based on exposure, testing, and both household and student symptoms.

The challenges tracker assesses the household for challenges experienced (listed in Figure 2C) in the last 14 days (based on those included in the PhenX Measures of the Social Determinants of Health¹), whether or not they have received assistance for those challenges, and whether they would like a family navigator to reach out to them to provide assistance with their reported challenges.

Figure 1. Flow Diagram for the mHealth Program
Every interaction made within the tool is tagged and tracked, along with the registered household that used it, the date, time, and location of each session. This data is written to a FERPA/HIPAA compliant database upon which refreshable dashboards for both the after-school program administrators and the family navigators are derived. The program administrator dashboard includes summaries by grade and school of symptoms experienced, whether households were advised to keep their student(s) home, reasons for the stay-at-home advisement, and challenges experienced. They also include named lists of students registered and day by day screening results. The navigator dashboard includes a list of households wishing to be contacted and their reported challenges.

For this pilot, a community navigator will be embedded within each after school program and households will receive gift card incentives upon registration and again upon reaching one month of weekly use sessions. Navigators will assist households with the process of finding and applying for aid for their specific challenges when possible. Some challenges are larger system issues which we are collecting to better advocate for future changes. This is made clear to households on each of the final screens within the challenges tracker. These navigators have worked in other roles with these programs before and are familiar with their households. Although not formally assessed for or tracked, the after-school programs had a history and reputation for assisting their families with challenges when they are confided in about them. The navigators will document all contact attempts and their results.

Results

We deployed the symptom screener alone in 2 public districts and 2 private schools in fall 2020, achieving an average of 80% registration and a median weekly screening rate of 40% in the first 8 weeks post deployment. The after-school program pilot went live on June 1st, 2021 and enrolled 46 students (of 200 enrolled for the summer) in 28 households within the first 2 weeks. Demographics of student tool registrations will be compared to those of the after-school program enrollees to determine whether those who register are representative. Proportion of students enrolled whose households used the symptom checker will be tracked on a daily and weekly basis. Proportion of students enrolled whose households used the challenges tracker, who report any of the challenges, and specific challenges reported will be tracked on a bi-weekly and monthly basis. Navigator requests, time to contact, and result of contact will be assessed weekly. Finally, participating households, after school program administrators, and family navigators will be interviewed periodically throughout the pilot to provide feedback on their experience with the program, the app, and the family navigation process.

Conclusion

The proposed program has the potential to enhance the ability of school-based programs to operate safely during the pandemic and to address the social and health disparities that arise. Study results will determine the feasibility and acceptability of such a program. If found to be feasible and acceptable, the program will provide scalable resources for deployment in future pandemics, natural disasters, and localized hardships.

References

Improving the Prediction of Healthcare Costs and Utilization Using the Area Deprivation Index and Statewide Insurance Data

Hadi Kharrazi, MD, PhD1, Hsien-Yen Chang, PhD1,2, Elham Hafez, MD, PhD1, Xiaomeng Ma, MS1, Jonathan P Weiner, DrPH1

1 Center for Population Health Information Technology, Johns Hopkins University, Baltimore, MD  
2 Center for Drug Safety and Effectiveness, Johns Hopkins University, Baltimore, MD

Introduction

Traditionally, risk adjustment models have relied on individual-level variables to predict and explain healthcare costs and utilization. However, as the performance of risk adjustment models fails to gain significant improvements with additional individual-level variables, utilizing aggregated information is increasingly being proposed as an alternative method to improve risk stratification and population health management efforts. Place-based determinants of health (i.e., the characteristics of the neighborhoods of patients’ residence) are powerful drivers of morbidity, mortality, and future well-being, yet they mostly remain outside the conventional medical care delivery systems. Adopting Area Deprivation Index (ADI) to represent place-based determinants, we aimed to evaluate its added values on claims-based risk adjustment models in predicting a range of utilization and cost outcomes.

Methods

This was a retrospective cohort study using 2013 and 2014 Maryland Medical Care Database (MMCD) from the Maryland Health Care Commission (MHCC). From 8,403,458 enrollees that had any record in MMCD between 2012 and 2016, we applied the following inclusion criteria: (1) aged 18-63 in 2013 with known sex; (2) Maryland residence in both 2013 and 2014; (3) 6+ month medical and pharmacy enrollment in both 2016 and 2017; and, (4) living in ZIP codes with 1000+ residents. Eventually, study population had 1,150,984 eligible enrollees. An individual’s ZIP code was assigned based on the most frequently occurring ZIP code and then the most recent ZIP code based on enrollment. The Area Deprivation Index (ADI) was adopted to represent place-based determinant of health, which is a validated measure of community disadvantage calculated at multiple geographical levels. ADI allows for rankings of neighborhoods by the socioeconomic disadvantage in a region of interest. ADI includes factors for the domains of income, education, employment, and housing condition; it can be used to inform health delivery and policy, especially for patients in the most disadvantaged neighborhoods. A percentile based on the national ADI ranking was assigned to each ZIP code and then attached to an individual through the geographic linkage. We generated three types of costs and five utilization markers in the concurrent (2013) and prospective (2014) calendar year at the individual level. Three types of costs were total costs, medical costs (non-pharmacy costs) and pharmacy costs; all costs were truncated at the respective top .5% of all positive costs. Five utilization markers included being top 5% top users, having any hospitalization, having any emergency department (ER) visit, having any avoidable ER visit, and having any readmission within 30-day of discharge. Control variables included two demographic variables (i.e., age and sex) and one morbidity measure. Age was divided into the following five categories: 18-24, 25-34, 35-44, 45-54, and 55-63 years old. Age was restricted to 63 in 2013 since people aged 64 in 2013 would enter Medicare in 2014 and we would not have their claims information. The morbidity measure was the DxRx-PM score, derived from the Johns Hopkins Adjusted Clinical Group (ACG) system. The DxRx-PM score is a comprehensive diagnosis-based predicted score constructed from various ICD and NDC-based indicators. The DxRx-PM score has been demonstrated to be a valid measure of morbidity. We applied linear regression for costs and logistic regression for binary utilization markers. We constructed two base models: one with demographics only and another with demographics and DxRx-PM score; for each base model, we constructed three ADI-enhanced models. Three sets of ADI variables entered separately to the base models: (1) ADI national percentile as a continuous variable; (2) ADI national decile as a categorical variable; and, (3) five components of ADI index, which contribute to more than 90% of variation of the ADI (i.e., median family income, home value, gross rent, monthly mortgage, and income disparity). The performance measures included R² from linear regression and AUC from logistic regression.

Results

Among 1.15 million eligible study subjects, 47.55% were male and the mean age was 41.2 years old. The average ADI national rank percentile was 47.3, suggesting that on average the sample lived in an area slightly more advantageous than the average American. Only 8.73% lived in the bottom 10% socioeconomic areas while 10.69% lived in the top 10% socioeconomic areas. The average total healthcare costs in both years were about $2,500, among which $1,600 were medical costs while $900 were pharmacy costs. About 2.7% had at least one hospitalization in
either year, about 12% had at least one ER visit, 7.7% had at least one avoidable ER visit, and .27% had at least one 30-day readmission. Adding ADI national rank percentiles (continuous variable), adding ADI national rank deciles (categorical variable) or adding the five components of ADI in either base model had minimum impact on the $R^2$ of total, medical or pharmacy costs (Table 1). Across five binary utilization markers, adding ADI only had the impact on two ER-related utilization markers (Table 2). For example, adding ADI national rank deciles to base model 2 statistically significantly increased AUC of predicting any avoidable prospective ER visit from .610 (95% CIs: .609-.612) to .613 (95% CIs: .611-.15), while doing so had no impact on predicting any prospective hospitalization (AUC of base model 2: .729, 95% CIs .726-.732; AUC of the base model 2 with ADI: .729, 95% CIs .726-.732).

### Table 1. Impact of adding ADI variables on $R^2$ across three levels on claims-based risk adjustment models in predicting healthcare costs

<table>
<thead>
<tr>
<th>Year</th>
<th>Outcome</th>
<th>2013 (Concurrent)</th>
<th>2014 (Prospective)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Total Cost</td>
<td>Medical Cost</td>
</tr>
<tr>
<td>Base 1: Age, Sex</td>
<td>.035 (.034-.035)</td>
<td>.036 (.035-.037)</td>
<td>.018 (.017-.018)</td>
</tr>
<tr>
<td>+ ADI US Rank Continuous</td>
<td>.035 (.034-.036)</td>
<td>.036 (.036-.037)</td>
<td>.018 (.017-.018)</td>
</tr>
<tr>
<td>+ ADI US Rank Deciles</td>
<td>.035 (.034-.036)</td>
<td>.036 (.036-.037)</td>
<td>.018 (.017-.018)</td>
</tr>
<tr>
<td>+ Five ADI Components</td>
<td>.035 (.034-.036)</td>
<td>.036 (.035-.037)</td>
<td>.018 (.017-.018)</td>
</tr>
<tr>
<td>Base 2: Age, Sex, Risk score</td>
<td>.643 (.639-.648)</td>
<td>.452 (.448-.456)</td>
<td>.377 (.372-.381)</td>
</tr>
<tr>
<td>+ ADI US Rank Continuous</td>
<td>.643 (.639-.648)</td>
<td>.452 (.449-.456)</td>
<td>.377 (.372-.381)</td>
</tr>
<tr>
<td>+ ADI US Rank Deciles</td>
<td>.643 (.640-.648)</td>
<td>.452 (.449-.456)</td>
<td>.377 (.372-.381)</td>
</tr>
<tr>
<td>+ Five ADI Components</td>
<td>.643 (.639-.648)</td>
<td>.452 (.449-.456)</td>
<td>.377 (.372-.381)</td>
</tr>
</tbody>
</table>

### Table 2. Impact of adding ADI variables on AUC across three levels on claims-based risk adjustment models in predicting healthcare utilization

<table>
<thead>
<tr>
<th>Year</th>
<th>Outcome</th>
<th>2013 (Concurrent)</th>
<th>2014 (Prospective)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Any ED</td>
<td>Any Avoidable ED</td>
</tr>
<tr>
<td>Base 1: Age, Sex</td>
<td>.549 (.547-.550)</td>
<td>.572 (.570-.573)</td>
<td>.545 (.544-.547)</td>
</tr>
<tr>
<td>+ ADI US Rank Continuous</td>
<td>.551 (.549-.552)</td>
<td>.575 (.573-.576)</td>
<td>.547 (.546-.549)</td>
</tr>
<tr>
<td>+ ADI US Rank Deciles</td>
<td>.554 (.552-.555)</td>
<td>.579 (.577-.580)</td>
<td>.550 (.549-.552)</td>
</tr>
<tr>
<td>+ Five ADI Components</td>
<td>.551 (.550-.553)</td>
<td>.576 (.574-.578)</td>
<td>.548 (.547-.550)</td>
</tr>
<tr>
<td>Base 2: Age, Sex, Risk score</td>
<td>.649 (.647-.650)</td>
<td>.654 (.653-.656)</td>
<td>.592 (.591-.594)</td>
</tr>
<tr>
<td>+ ADI US Rank Continuous</td>
<td>.648 (.646-.649)</td>
<td>.655 (.653-.656)</td>
<td>.593 (.591-.594)</td>
</tr>
<tr>
<td>+ ADI US Rank Deciles</td>
<td>.648 (.647-.650)</td>
<td>.656 (.655-.658)</td>
<td>.594 (.592-.595)</td>
</tr>
<tr>
<td>+ Five ADI Components</td>
<td>.648 (.646-.649)</td>
<td>.655 (.653-.657)</td>
<td>.593 (.591-.594)</td>
</tr>
</tbody>
</table>

### Discussion

Our study is one of the first to evaluate the impact of neighborhood characteristics on the performance of risk adjustment models. The only similar study we identified did not find neighborhood characteristics contributing to risk prediction beyond demographic (e.g., age and race) and clinical data in the EHR from a ~90k participants of a single county. Thus, our study added considerably to the limited literature on the impact of incorporating place-based determinants of health into risk adjustment models. The only significant impact of adding ADI on the performance of risk adjustment models was on binary indicators of an ER visit, especially an avoidable ER visit. It can be hypothesized that patients living in neighborhoods with lower SES were less likely to be able to afford seeking healthcare under the regular setting given the cost burden and the lack of healthcare providers; therefore, they were more likely to use ER as their regular source of healthcare.

### Conclusion

Adding ADI to the claims-based risk adjustment models improves the ability of risk adjustment models in predicting the probability of having any ER visit or having an avoidable ER visit. Future research should focus on patients with higher-need for social services (e.g., Medicaid patients), assess more granular place-based determinants (e.g., Census block group), and evaluate the added-value of individual social variables instead of a composite index such as ADI.

### References

FHIRedApp: A LEAP in Health Information Technology for Promoting Patient Access and Health Equity
Anjum Khurshid, MD, PhD, Eliel C. de Oliveira, MS, MBA, Vidya Lakshminarayanan, MS, Vishal Abrol, BS.
Dell Medical School, The University of Texas at Austin, Austin, TX

Introduction
Nearly a year ago, the Office of the National Coordinator for Health Information Technology (ONC) and the Centers for Medicare & Medicaid Services (CMS) released final regulations to implement the information blocking and patient access provisions of the 21st Century Cures Act. These regulations underscore the United States government’s recognition that healthcare information technology has not kept pace with other industries when it comes to the adoption and utilization of modern interoperable technologies and in empowering consumers.

We developed the FHIRedApp as an App Platform that allows patients to gain access to their medical data and share that access as HL7® FHIR® (Fast Healthcare Interoperability Resources) APIs with third party App developers. This allows for innovative solutions to be built and provided to patients to harness the value of their medical data. This effort was funded by ONC in 2019, as part of ONC’s Leading Edge Acceleration Projects (LEAP), to design, develop and demonstrate an enhanced PET platform. The goal of the ONC program was to modernize patient engagement tools, engage diverse populations including the underserved, enhance opportunities for underrepresented populations to participate in research while addressing security and privacy concerns, allowing data transport from various digital devices, and applying user-centered design principles.

Methods

Design through Community Engagement Studios and Human-Centered Design
The overall development of FHIRedApp was informed by Community Engagement Studio (CES) and Human-Centered Design (HCD) methodologies. CES is a structured-forum, designed to gain valuable community insights on a research topic. (2) CES provides researchers with a real-time, in-person opportunity to share their research projects and goals with community members, advocates, and other stakeholders, and to receive direct feedback and guidance via a facilitated conversation. CES meetings are led by a community navigator. We ran CES sessions engaging community experts from the local African American, Asian American and the Latinx communities.

The insights gained from CES sessions informed the HCD work to uncover the features and capabilities that underserved community desired from an App platform such as FHIRedApp. The HCD methodology included: a) a competitive analysis of current similar platforms; b) scenario creation based on input from CES meetings and surveys; c) development of early low-fidelity designs; d) usability tests; and e) iterative one-on-one design sessions with users to develop high-fidelity designs.

Results

Figure 1 depicts a flow diagram on how the FHIRedApp back-end infrastructure is able to aggregate clinical data and seamlessly deliver such data as APIs to users of the FHIRedApp Platform. We partnered with the Integrated Care Collaboration (ICC), an HIE in Central Texas, that aggregates data of over a million patients across various healthcare providers in the region. ICC data contributors include three hospital systems accounting for about 80% of all hospital services in the region, federally qualified health centers, behavioral health providers, the Austin-Travis County Emergency Medical Services (EMS), and several primary care provider organizations sharing Medicare, Medicaid and uninsured patients’ data across all participants in the exchange.

Additionally, we adopted a Privacy-Preserving Record Linkage (PPRL) solution that is utilized by several large national data research networks such as PCORnet to link patient records without storing Personally Identifiable Information (PHI) such as names, addresses, or phone numbers and enhance the security and privacy of the data.

The process of data integration and linkage for FHIRedApp back-end infrastructure happens on Amazon Web Service (AWS) accounts managed by our team through the following steps:

1. ICC deploys a Hashing Tool that accesses demographics information (names, date of birth, gender, etc.) to generate salted hashes that comply with HIPAA’s de-identification methods and a cross-table of IDs between the hashes generated and the CDR. The salted hashes are then submitted to our team along with a limited dataset as defined by HIPAA which excludes most PII elements except for dates and zip codes through Extract, Transform and Load (ETL) processes;

2. when a user signs up and consents to use the platform, FHIRedApp collects basic demographic information from the user (name, email, phone number, and gender) that will be used by the PPRL solution (hashing and matching tools) to link users to their medical records. The hashing tool deployed at the back-end receives the user’s demographic information and also generates salted hashes and a cross-table of IDs that links users to their hashes;
3. Finally, an ETL process utilizing the IDs linked by the Matching Tool combines the records from the LDS and the users’ database (DB) to generate the FHIRedApp DB. That database is transformed into FHIR APIs and exposed using a HAPI Server to the mobile FHIRedApp front-end App.

Hash function is any function that can be used to map digital data of arbitrary size to digital data of fixed size, with slight differences in input data producing very big differences in output data. The unique aspects of using a hash function are due to the following properties:

1. Each hash is unique but always repeatable - that means that a word such as 'car' will hash to something that no other word hashes to, but it will always hash to the same thing;

2. The function is 'one way' - meaning that if you are given the value of what 'car' hashes to but you didn't know what made it, you would never be able to find out that 'car' was the original word.

Discussion
We describe the design and development of FHIRedApp, an App platform that allows patients to access their data and to share that access as HL7® FHIR® (Fast Healthcare Interoperability Resources) APIs with third party App developers. As a Leading Edge Acceleration Project for health information technology sponsored by ONC, we are accomplishing two major tasks: first, to demonstrate the use of interoperability and authentication standards approved by ONC and the industry, such as HL7® FHIR and OAuth2, to help develop responsive patient engagement technologies, and second, to co-develop and co-design FHIRedApp with active involvement of African-American, Latinx, and Asian-American community members to help address health disparities in the use and benefits of PETs in our populations. We hope that the methodology and design of FHIRedApp will facilitate further development of these approaches and meaningful engagement of vulnerable populations in fulfilling the mandate of 21st Century Cures Act to allow patient to “access, exchange, and use” their health data “without special efforts”.

Reference


Clinical Concept Extraction Using Contextual String Embeddings

Youngjun Kim, PhD\(^1\), Paul M. Heider, PhD\(^2\), Stéphane M. Meystre, MD, PhD\(^2\)
\(^1\)Sema4, a Mount Sinai venture, Stamford, Connecticut, USA
\(^2\)Medical University of South Carolina, Charleston, South Carolina, USA

**Introduction:** Identifying mentions of concepts in electronic health record (EHR) narrative text is a fundamental problem and its importance is still growing as a prerequisite for other high-level tasks. Clinical concept extraction has often been framed as a sequential labeling problem to assign a semantic category label to each word in a sequence. Deep learning-based algorithms with pre-trained word embeddings\(^1\) have provided a robust solution. Recently, context-dependent embeddings\(^2-4\) have been used to construct vector representations of words as input to neural networks. Several studies leveraging these embeddings have demonstrated improved accuracy for concept extraction from unstructured clinical narratives.\(^5,7\) Our research aims to improve clinical concept extraction using syntactic and semantic knowledge acquired from large amounts of unlabeled clinical text. We employ contextual *string* embeddings called *Flair*\(^3\) to transfer that knowledge into our target tasks. These embeddings learn a word as a sequence of characters and generate the word representation differently depending on its context. Although Flair embeddings have achieved state-of-the-art performance in general text NER (named entity recognition)\(^1\), few studies\(^5,9\) have used them in clinical NLP. If compared to BERT\(^4\), a popular transformer-based word embeddings resource applied in clinical NLP, Flair offers a convenient way to handle longer sequences of words. To our knowledge, this is the first work that applies Flair embeddings for de-identification of EHR narratives.

**Methods:** We focus on three representative clinical natural language processing (NLP) tasks. The first task was defined for the 2010 i2b2 (integrating biology and the bedside) challenge\(^10\) and involved extracting three categories of medical concepts: Problem, Treatment, and Test. The next task is text de-identification introduced for the 2014 i2b2 challenge\(^11\). It aims to detect pre-defined categories of personally identifiable information (PII). These categories include Name, Profession, Location, Age, Date, Contact, and ID. The final task was introduced in the 2018 national NLP clinical challenges (n2c2) shared task\(^12\) and focuses on extraction of medications and related information (e.g., dosage, route, frequency) from clinical text.

In this study, we pre-trained two versions of the language model from clinical text. One was trained using fastText\(^1\) as static word embeddings and the other was trained using the Flair\(^3\) software library. We used clinical notes from the MIMIC-III clinical database\(^13\) v1.4 (we call it ‘MIMIC embeddings’). Models were trained with all 2,083,180 clinical notes. We created a 300-dimensional skip-gram fastText model. The learning rate and number of epochs were set to 0.02 and 10, respectively. Flair embeddings were trained on the same corpus. We set the number of hidden states of the LSTM (long short-term memory) to 2048 and trained the Flair model for 10 epochs. Pre-training was performed on a NVIDIA Tesla P4 GPU and took about 1 month. For a comparative assessment with our MIMIC word embeddings, the publicly available fastText\(^1\) and Flair\(^3\) models were used as general text embeddings.

We used the bidirectional LSTM (Bi-LSTM) algorithm to solve the structured prediction problems. A CRF (conditional random fields) classifier was added as the last layer in the network. We created four different variations of LSTM-CRF models for each task: 1) fastText with general text embeddings, 2) fastText with MIMIC embeddings, 3) combined fastText and Flair vectors with general text embeddings, and 4) fastText and Flair with MIMIC embeddings. We set the learning rate to 0.001 and trained the Bi-LSTM models for 10 epochs with 50% dropout. We calculated the average value between five trials of each model because of its non-deterministic results. Instead of ignoring the individual trial outputs, we created a voting ensemble method that combined these trial predictions to further improve performance. The voting ensemble selected all concepts that received at least three votes.

**Results:** We used the official evaluation script provided for each shared task to calculate primary performance metrics. For the 2010 and 2014 i2b2 shared tasks, a true positive is obtained when both the text span and concept category exactly match the reference annotation. For the 2018 n2c2 shared task, a match is counted if there is any overlap between the reference standard text span and the concept detected by the system. Table 1 shows the results produced with combinations of NLP tasks, text sources of pre-trained models, and embedding approaches. The impact of contextual embeddings varies across the tasks (when comparing rows 1 and 2). Adding Flair representations allows for substantial performance gains for the 2010 i2b2 task. This implies that contextual embeddings can more effectively handle medical concepts involving long modifiers or unseen words. The F\(_1\)-scores of the general text and MIMIC embeddings increased from 85.4% and 86.4% to 86.2% and 88.6% respectively. In the 2014 i2b2 task, Flair embeddings increased the F\(_1\)-scores between 0.6 and 0.8%. In the 2018 n2c2 task, Flair embeddings further benefited
concept categories that were recorded in a more unstructured manner with rich terminologies, such as reason (for prescription) and adverse drug event (ADE). We created BERT models with the BlueBERT-Large version14 for comparison. BERT embeddings produced a slightly higher F1-score (93.6%) than Flair embeddings in the 2018 n2c2 task. Detailed comparisons will be provided in a more extended publication.

MIMIC embeddings allowed for an improvement over their counterpart general text embeddings (when comparing the column “General text” with the column “MIMIC text”). Both fastText and Flair embedding pre-trained with the MIMIC corpus improved performance. For the 2010 i2b2 and 2018 n2c2 tasks, MIMIC embeddings increased the F1-scores between 0.4 and 2.4%. As with the results reported by Alsentzer et al.6, text de-identification was an exception, presumably because PHI phrases in MIMIC texts were re-synthesized with tags describing the type of information (e.g., “[ Name ]” instead of “Mary Jones”). We also confirmed that each voting ensemble consisting of multiple trials outperformed the individual classifiers. The voting ensemble increased the F1-scores between 0.3 and 1.2%. The best result for each task (bold in Table 1) was significantly better than all other methods at the 95% significance level.

**Table 1.** Accuracy results of each clinical concept extraction models.

<table>
<thead>
<tr>
<th>Method</th>
<th>2010 i2b2 medical concepts</th>
<th>2014 i2b2 de-identification</th>
<th>2018 n2c2 ADEs and medications</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>General text</td>
<td>MIMIC text</td>
<td>General text</td>
</tr>
<tr>
<td>fastText</td>
<td>Pre</td>
<td>Rec</td>
<td>F1</td>
</tr>
<tr>
<td></td>
<td>86.1</td>
<td>84.7</td>
<td>85.4</td>
</tr>
<tr>
<td>Flair</td>
<td>86.8</td>
<td>85.5</td>
<td>86.2</td>
</tr>
<tr>
<td>Vote(fastText)</td>
<td>87.9</td>
<td>85.3</td>
<td>86.6</td>
</tr>
<tr>
<td>Vote(Flair)</td>
<td>88.6</td>
<td>86.3</td>
<td>87.4</td>
</tr>
</tbody>
</table>

Pre: precision, Rec: recall, F1: F-score, Vote(X): voting ensemble of five trials of X

**Conclusion:** We demonstrated that contextualized string embeddings, especially trained with text resources that share similar syntactic and semantic word distributions, could be beneficial to clinical concept extraction. We confirmed that vector representations learned from such corpora (e.g., MIMIC-III clinical database) can more accurately detect concept boundaries and identify unseen concepts from the training data. To advance our neural network architectures, we will augment the vector representation using ELMo2 and BERT4 context embeddings.

**Acknowledgments:** Partly supported by the National Institute for General Medical Sciences (R42GM116479).

**References**

Feasibility and efficacy of using commercially available mHealth technologies and tailored text messages to improve self-care in patients with heart failure: Phase I clinical trial of the iCardia4HF intervention

Spyros Kitsiou, PhD1, Susan J. Pressler PhD, RN2, Mayank M. Kansal, MD1,4, Susan W. Buchholz, PhD, RN3, Jinsong Chen, PhD, Jonathan Leigh, MPH, Ayomide J. Owoyemi, MBChB, MPH1, Ben S. Gerber, MD, MPH5

1University of Illinois at Chicago, Chicago, IL; 2Indiana University, Indianapolis, IN; 3Michigan State University, East Lansing, MI; 4Jesse Brown VA Medical Center, Chicago, IL; University of Massachusetts Medical School, Worcester, MA

Introduction

Heart failure (HF) is a chronic condition associated with substantial morbidity, mortality, and hospital readmissions1. Evidence shows that effective HF self-care is crucial to improving HF outcomes2. Self-care is a naturalistic decision-making process that involves self-care maintenance (taking medications as prescribed, restricting sodium intake, and staying physically active), symptom perception (monitoring and interpreting HF signs and symptoms), and self-care management (response to symptoms when they occur)3. Self-care is poor among HF patients, especially older adults and underserved ethnic minorities due to lack of skills, confidence, knowledge about HF, and health beliefs2. There is an urgent need for low-cost and scalable interventions to assist patients with HF self-care at home. The objective of this study was to assess the feasibility, acceptance, and effects of a patient-centered intervention (iCardia4HF) that uses commercially available mobile health (mHealth) technologies and text-messages (TM) to improve HF self-care.

Methods

Study design: A prospective randomized controlled trial was conducted between January 2019 and February 2020 (trial registration: NCT03642275). Study participants were actively recruited from the University of Illinois Health Sciences and Hospital System (UI Health). Key research personnel used the electronic medical record system to screen patients admitted to the hospital due to HF and outpatients with upcoming appointments to the HF clinic. Potentially eligible patients were approach for further screening and consent. Randomization was performed after baseline assessment, using QMinim (a free online minimization software) to ensure better-than-chance group balance in terms of age, gender, and HF severity based on the New York Heart Association (NYHA) class as in previous trials5.

Eligibility criteria: Patients were eligible to enroll in the study if they had Stage C HF (i.e., current or prior HF symptoms), were ≥40 years of age, spoke English, owned a smartphone, and lived within a 30 mile-radius from UI Health. Patients were excluded if they had end-stage renal disease/dialysis, end-stage HF, active cancer, cognitive impairment (Montreal Cognitive Assessment score of <22), were unable to perform self-care, or lived in nursing home.

Control group: Participants in the control group received usual care consisting of patient education (literacy-sensitive material about HF self-care developed by the Sheps Center for Health Services Research) before hospital discharge and follow-up visits at the outpatient HF clinic 1-2 weeks after discharge and 1-3 months thereafter.

Intervention: In addition to usual care, participants in the intervention group received the iCardia4HF intervention which comprises: (1) the Heart Failure Storylines mobile app; (2) three connected health devices (Withings Body Cardio Scale and Blood Pressure Monitor, and Fitbit Charge 2 activity tracker) along with their respective apps (Health Mate and Fitbit); and (3) three individually tailored TMs per week targeting health beliefs, self-efficacy, and HF-knowledge. HF Storylines is a free, web-based app developed by Self Care Catalysts in collaboration with the Heart Failure Society of America. It interfaces with the Withings and Fitbit devices and provides patients with a number of tools to support self-monitoring of HF sing and symptoms, including medication adherence and low-sodium diet.

Outcomes: The primary outcome was change in HF self-care including self-efficacy, measured with the Self-care Heart Failure Index (SCHFI v6.2)3 at 30 and 60 days. Outcome assessors were blinded to the group allocation. Feasibility was assessed in the domains of recruitment and retention. Intervention acceptance was assessed with the Technology Acceptance Model3 and semi-structured interviews post-intervention. Adherence to daily self-monitoring was objectively assessed using time-stamped data transmitted from the mobile apps to our study server via Web APIs.

Study procedures: Intervention participants received 1-hour training at baseline along with printed instructions on how to use the devices and mobile apps. Daily self-monitoring included weight and body composition measures (e.g., fat mass, water mass, and muscle mass) with the Withings Body Cardio scale, blood pressure (BP) and heart rate with the Withings BP cuff, and physical activity/inactivity measures (e.g., steps, intensity of activity, sedentary minutes, continuous heart rate, and sleep) captured by the Fitbit activity tracker.
Results

Twenty-seven patients were enrolled and randomly assigned to the intervention (n=13) or control group (n=14). One patient from the intervention group was lost to follow-up, while a second one was admitted to a nursing home and became ineligible before allocation of the mHealth apps and devices. Thus, a total of 25 patients (mean age 56, standard deviation [SD] ±8.3, 44% female) were included in the full analysis set (11 intervention, 14 control).

Baseline characteristics (age, gender, race, education, income, weight, BP, and HF severity) were similar between the two groups. Most patients were Black (91%) with no more than high school education (73%), functionally compromised (76% had NYHA Class II or III), and obese (mean weight: 110.8 kg, SD ±33.1; mean Body Mass Index: 36.6 kg/m², SD ±10.1). Mean left ventricular ejection fraction was 35.8% (SD ±12.4). Although not reaching statistical significance due to the small sample size, participants in the intervention group experienced greater improvements in HF self-care maintenance, symptom perception, and self-efficacy, compared to the control group over 60 days (Table 1). Effect sizes ranged from small (0.20) to moderate (0.50) based on Cohen’s D.

### Table 1: Self-care Heart Failure Index

<table>
<thead>
<tr>
<th>HF self-care domains</th>
<th>Control</th>
<th>Intervention</th>
<th>P1</th>
<th>P2</th>
<th>Cohen’s D (30day-baseline)</th>
<th>Cohen’s D (60day-baseline)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>D30day - baseline (SD)</td>
<td>D30day - baseline (SD)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Self-care maintenance</td>
<td>-3.75 (12.82)</td>
<td>3.75 (12.51)</td>
<td>6.14 (15.39)</td>
<td>6.36 (15.34)</td>
<td>0.229</td>
<td>0.765</td>
</tr>
<tr>
<td>Symptom perception</td>
<td>6.68 (24.03)</td>
<td>-4.55 (25.41)</td>
<td>5.92 (18.04)</td>
<td>9.36 (15.49)</td>
<td>0.323</td>
<td>0.066</td>
</tr>
<tr>
<td>Self-care management</td>
<td>-12.42 (20.28)</td>
<td>0.43 (13.44)</td>
<td>5.79 (17.85)</td>
<td>3.86 (24.55)</td>
<td>0.067</td>
<td>0.784</td>
</tr>
<tr>
<td>Self-care efficacy</td>
<td>-0.50 (8.48)</td>
<td>-2.50 (13.33)</td>
<td>5.68 (13.92)</td>
<td>7.50 (17.21)</td>
<td>0.265</td>
<td>0.166</td>
</tr>
</tbody>
</table>

Adherence to daily self-monitoring of weight and BP with the devices and app was high during the intervention period (mean ±SD: 89.8 ±7.04% and 83.3% ±6.29%, respectively), with no statistically significant decline over time. Participants adhered well to wearing the Fitbit tracker with the mean participant logging 600 minutes/day or more on 85.3% of intervention days (51/60). Mean Fitbit wear-time was 1031 ± 305 minutes/day during the intervention period. Participants perceived the iCardia4HF intervention to be useful (mean 2.94±0.07 on a 3-point Liker scale, where 1=disagree and 3=agree) and easy to use (2.97±0.08). Mean user satisfaction was 2.94±0.07.

Discussion

The results of this small RCT provide initial evidence that iCardia4HF might be a feasible and acceptable intervention for promoting HF self-care in patients with HF. Participants were satisfied with the intervention and found the multi-connected app kit easy to use. Minor technical challenges such as forgetting how to charge/sync the Fitbit tracker or receiving a measurement error message from the weight scale or BP monitor when not properly positioned were addressed in a timely manner (within 48 hours) via telephone. Findings provide support for the conduct of a larger trial with longer follow-up (currently underway with funding from the National Institute of Nursing Research R21NR018281, NCT 04262544) to determine the efficacy of iCardia4HF intervention. To our knowledge, this is the first RCT that evaluated the feasibility of using commercially available mHealth technologies to promote self-care in patients with HF. In a recent systematic review\(^2\), we found that previous RCTs have evaluated either research-based mHealth apps that are not commercially available or apps that are part of a larger remote patient monitoring system in which clinicians are actively monitoring the incoming data to provide patients with clinical feedback when needed. The iCardia4HF study represents a meaningful advance in standalone, patient-centered mHealth interventions, and an important step in identifying a scalable solution that has the potential to bring about a new paradigm in HF self-care. The combination of smartphones and commercially available mHealth technologies make the iCardia4HF intervention portable to different clinical settings and create exciting opportunities for scalability and broader impact.

References


Acknowledgement

This study was funded by the National Institute on Aging through grant P30AG022849.
Challenges and Solutions to Promoting Evaluation Practices in Software Development Process within an Academic Medical Center

Polina Kukhareva, PhD, MPH, MS; Charlene Weir, PhD, RN; Thomas Reese, PhD, PharmD; Teresa Taft, PhD; Guilherme Del Fiol, MD, PhD; Kensaku Kawamoto, MD, PhD, MHS

University of Utah, Salt Lake City, Utah

Introduction: Electronic health records (EHRs) are evolving into platforms for EHR-based apps, enabling an emerging EHR app ecosystem. With millions of patients affected by locally-developed EHR apps, it is critically important to rigorously evaluate these apps across all stages of their lifecycle. The software development process could be improved by applying concepts, methods, and evaluation strategies from human-centered design, implementation science, and health technology assessment (HTA) domains. We sought to demonstrate the feasibility of promoting standardized evaluation practices in local software development. University of Utah’s ReImagine EHR initiative developed and implemented into clinical practice several interoperable apps and was a perfect place for demonstrating the value of standardized evaluation approaches. The resulting EHR-based Digital Innovation Evaluation Framework (EDIEF) is described.

Methods: We aimed to develop and apply a pragmatic, mixed-methods, full-lifecycle evaluation framework for EHR-based apps. Drawing from human-centered design, implementation science and health technology assessment perspectives, we selected evaluation approaches that could be used at all EHR apps lifecycle phases: exploration, preparation, implementation and sustainment (1). The novel EDIEF was iteratively developed through literature review, expert consensus, and evaluating apps in real-world settings. We tested the feasibility of using EDIEF in the development of a representative app (a SMART on FHIR Neonatal Bilirubin App, which won the 2019 AMIA/Health Level Seven (HL7) FHIR App Showcase). We describe here the challenges encountered and our recommendations.

Results: We designed a new evaluation framework known as EDIEF and implemented it across the software development lifecycle (Figure 1). Examples of assessments for each evaluation step are provided in Table 1.

Table 1. EDIEF: Software Lifecycle Phases and Evaluation Steps

<table>
<thead>
<tr>
<th>Phase</th>
<th>Evaluation Steps</th>
<th>Examples of Assessments</th>
</tr>
</thead>
<tbody>
<tr>
<td>I. Exploration</td>
<td>(1) Business case</td>
<td>• Potential benefits, risks, costs, feasibility, equity, and alternatives</td>
</tr>
<tr>
<td></td>
<td>(2) Stakeholder requirements</td>
<td>• User needs, other stakeholder needs, and measures</td>
</tr>
<tr>
<td></td>
<td>(3) Technical requirements</td>
<td>• Technical portability and interoperability needs</td>
</tr>
<tr>
<td>II. Preparation</td>
<td>(4) User-centered design</td>
<td>• Usability, unanticipated user needs, and alternative uses</td>
</tr>
<tr>
<td></td>
<td>(5) Technical acceptability</td>
<td>• Portability, software performance, reliability, security, and privacy</td>
</tr>
<tr>
<td></td>
<td>(6) Comparative benefits</td>
<td>• Efficiency and efficacy as compared to usual care</td>
</tr>
<tr>
<td>III. Implementation</td>
<td>(7) Implementation planning</td>
<td>• Organizational characteristics, implementation resources</td>
</tr>
<tr>
<td></td>
<td>(8) Formative implementation</td>
<td>• App adoption, usage patterns, implementation strategy effectiveness</td>
</tr>
<tr>
<td></td>
<td>(9) User satisfaction</td>
<td>• System Usability Scale (SUS), interviews</td>
</tr>
<tr>
<td>IV. Sustainment</td>
<td>(10) Summative implementation</td>
<td>• App adaptations, technical performance, dissemination value</td>
</tr>
<tr>
<td></td>
<td>(11) Summative effectiveness</td>
<td>• Process outcomes, equity, targeted health outcomes (eg, weight)</td>
</tr>
<tr>
<td></td>
<td>(12) Economic evaluation</td>
<td>• Cost-effectiveness analysis, return on investment</td>
</tr>
</tbody>
</table>
To test and improve the EDIEF, we used all 12 evaluation steps in the evaluation of the Neonatal Bilirubin App including engagement of pediatrics in business case analysis, the application of usability heuristics principles, comparative efficiency testing, user adoption assessment, user satisfaction assessment, and the evaluation of process and clinical outcomes.(2) The use of the app resulted in 66 seconds saved per bilirubin assessment (95% CI, 53-79); the app was used for 90% of babies; system usability score for attending providers was 91 (95% CI, 86-96) which corresponds to “best imaginable” usability; clinically appropriate phototherapy during hospitalization increased for newborns with bilirubin levels above the guideline-recommended threshold (odds ratio, 1.84; 95% CI, 1.16-2.90; P = .01).(2) Additionally, we evaluated technical portability of the app software.(3)

As we utilized the EDIEF, we encountered challenges such as high cost involved in conducting rigorous evaluations, timeline pressures leading to skipping evaluation steps; and the need to evaluate a moving target. Table 2 outlines the key challenges encountered and our proposed solutions.

Table 2. Key Challenges and Proposed Solutions in Establishing an Evaluation Program

<table>
<thead>
<tr>
<th>Challenges</th>
<th>Proposed Solutions</th>
</tr>
</thead>
<tbody>
<tr>
<td>App evaluation requires significant expertise and resources.</td>
<td>•Recognize the importance and complexity of rigorous evaluation and invest accordingly in an evaluation team</td>
</tr>
<tr>
<td>•High cost of app evaluations</td>
<td>•Design apps using standards (e.g., SMART on FHIR, HL7 Clinical Decision Support (CDS) Hooks services, HL7 Clinical Quality Language rules) that can be re-used for automated evaluation</td>
</tr>
<tr>
<td>•Regulatory burden (e.g., institutional review board (IRB) and clinicaltrials.gov registration)</td>
<td>•Use a systematic and efficient approach to app evaluation</td>
</tr>
<tr>
<td>•Complexity of the healthcare domain</td>
<td>•Develop re-usable evaluation resources such as IRB application templates and interview guides</td>
</tr>
<tr>
<td>•Scarcity of multi-disciplinary evaluation experts</td>
<td>•Recognize the importance and complexity of rigorous evaluation and invest accordingly in an evaluation team</td>
</tr>
<tr>
<td>•Systematic under-funding of app evaluation as compared to development funding</td>
<td>•Design apps using standards (e.g., SMART on FHIR, HL7 Clinical Decision Support (CDS) Hooks services, HL7 Clinical Quality Language rules) that can be re-used for automated evaluation</td>
</tr>
<tr>
<td>Skipping evaluation steps can lead to problems in later project phases.</td>
<td>•Plan for, and insist on, adequate requirements gathering, testing and pilot implementation prior to wider release</td>
</tr>
<tr>
<td>•Failure to translate results of the early requirements analysis to design and evaluation</td>
<td>•Ensure that clinical champions are on board and available to provide guidance and feedback throughout the project phases. Consulting clinical champions can provide valuable feedback at a reasonable cost</td>
</tr>
<tr>
<td>•Timeline pressures leading to inadequate functional requirements gathering and app testing</td>
<td>•Rather than skipping an evaluation step entirely, conduct a pragmatic, lower-cost evaluation</td>
</tr>
<tr>
<td>•Difficulties enrolling users in testing programs, busy schedules of provider champions</td>
<td>•Recognize that multiple assessments may be required as the app and context evolve.</td>
</tr>
<tr>
<td>Need to evaluate a moving target</td>
<td>•Keep good documentation of when and how the app evolved following initial clinical deployment.</td>
</tr>
<tr>
<td>•Dynamic nature of health care, clinical guidelines, the EHR, stakeholder needs, and apps.</td>
<td>•Plan for the intervention to evolve.</td>
</tr>
<tr>
<td>•Summative evaluation methods not designed for evaluating a moving target</td>
<td>•Recognize that multiple assessments may be required as the app and context evolve.</td>
</tr>
</tbody>
</table>

Discussion: Inadequate evaluation during the development of EHR-based apps could lead to disjointed EHR user experiences with the final products, wasted resources, and unintended consequences including patient harm. As EHR-based apps become increasingly important in healthcare systems’ strategies to support clinical decisions, a pragmatic and comprehensive evaluation approach is critically needed. The EDIEF described here could facilitate these needed evaluations of emerging apps.

Conclusion: Commitment to systematic evaluation is critical to ensure the emerging EHR app ecosystem results in intended outcomes. This study demonstrates feasibility of promoting evaluation practices in local software development in medical centers. The evaluation framework described was successfully used to evaluate a number of locally developed EHR-based apps. Further research is needed to test and refine this pragmatic framework for evaluating EHR-based apps.

References:
Improving MHC Class I Antigen Processing Prediction via Representation Learning and Cleavage Site-Specific Kernels

Patrick J. Lawrence¹, Xia Ning PhD¹,²,³

¹Biomedical Informatics Department, The Ohio State University, Columbus, OH, US; ²Computer Science and Engineering Department, The Ohio State University, Columbus, OH, US; ³Translational Data Analytics Institute, The Ohio State University, Columbus, OH, US

1 Introduction
The major histocompatibility complex (MHC) Class I protein is a vital part of the immune system’s response to intracellular invasion by viruses, bacteria, parasite and tumorogenesis [1]. Its primary responsibility is to present antigens – typically, short peptides cleaved from proteins and are only 8-10 amino acids in length – into the extracellular environment to be recognized by cytotoxic (CD⁸⁺) T cells, which subsequently eliminate compromised cells via apoptosis. Thus, these peptides can be leveraged for the development of both vaccines that prime CD⁸⁺ T cells against a pathogen and drugs that elicit cytolytic activity in tumor cells. As a result, computer-aided methods have been developed to identify candidate peptides [1]. Among them, methods that rank peptides’ binding affinities to MHC class I molecule(s) have achieved the superior performance [2]. However, these methods may biological relevance; there is no guarantee highly ranked peptides will be selected for presentation by upstream proteins [3]. Recent attempts have been made to produce models that rank the likelihood of peptides being processed within the MHC class I presentation pathway [3][2]. In this work, we propose a novel antigen processing prediction model: mhcRank. Based on the architecture used by O’Donnell et al [3], our model imparts additional biological relevance, focusing on the carboxyl (C)-terminal cleavage site of the antigen and pre-processing antigen sequences to simulate what is observed in vivo [3]. Our model, mhcRank, also departs from the widely used BLOSUM62 matrix for amino acid representations in favor of learning problem-specific embeddings for each amino acid. Our experiments over benchmark data set demonstrate that mhcRank achieves significant performance improvement over the gold standard methods netMHCpan-4.0 eluted ligand and MHCflurry-2.0 antigen processing, referred to as netMHCpan and MHCflurry, respectively.

2 Methodology
The model architecture of our mhcRank is presented in Figure 1. In mhcRank, peptides are first processed to a uniform length via the process outlined in Figure 2. Processed peptides and flanks are represented using amino acid-specific learned embeddings. The proposed mhcRank has multiple components, each containing a series of 1-D convolutional layers aimed at (1) determining the probability of C and N terminal break at both the cleavage sites and within the peptide itself, (2) identifying characteristics of both flanking regions, and (3) capturing the C-terminus cleavage site. Components (1) and (2) are similar to those employed by MHCflurry [3]. Component (3), the cleavage site-specific kernel (CSSK), utilizes a global kernel to capture a representation of the entire cleavage site (purple box on right of Figure 1). Results from the convolutional layers and the peptide’s original length (light blue box on left of Figure 1) are all concatenated and enter two fully-connected layers before reaching the output layer. Models were trained identically to MHCflurry-2.0, including the dataset and training process [3]. This involved a binary cross-entropy loss function and Adam optimizer over maximum 500 epochs with a patience of 30 epochs. Post training, models were selected for inclusion in ensembles based on their respective AUC values on validation data. Mean precision and NDCG at k (with a 95% CI) were obtained using bootstrap resampling. Significance was determined by whether or not the CIs overlapped. These metrics were selected to evaluate how well mhcRank ranks peptides compared to the baseline methods. The values of k selected (10, 25, 50, 100, 250, 500) were those which might be realistically used in drug and vaccine development.
A. Peptides shorter than the desired length are padded with an ambiguous amino acid 'X' to the center. Raw peptides with an odd length have padding offset from the center by one amino acid towards the N-flank.

B. Peptides longer than the desired length are trimmed from the center.

C. Trimming to odd-length peptides requires the trim to be offset from the center by one amino acid towards the N-flank.

Figure 2: Peptide pre-processing

3 Experimental Results

Table 1 highlights the percent improvement in performance obtained by mhcRank compared to both netMHCpan and MHCflurry. The model mhcRank demonstrated statistically significant improvement over MHCflurry-2.0 for both metrics and at all values of $k$. Specifically, mhcRank had a mean improvement for all examined values of $k$ of about 9.83% and 10.12% over MHCflurry for precision and NDCG, respectively. The performance of mhcRank was much more comparable to the netMHCpan baseline with a mean percent improvement in precision and NDCG of roughly 1.79% and 2.35%, respectively. Even so, mhcRank still outperforms netMHCpan for most values of $k$. While not displayed here, it should be noted that the percent improvement increases further for both metrics as the values of $k$ increase, reaching a maximum of roughly 12-15% boost on both baseline methods.

4 Discussion

Peptide representations are enhanced through both pre-processing and learned embeddings. First, during pre-processing, we leverage biological knowledge to retain only the most relevant information. The dearth of enrichment in the central amino acids of the top 100 ranked peptides as shown in Figure 3 (generated using Seq2Logo [4]), paired with our improved performance, confirms the central amino acids are irrelevant features, whose removal promotes model generalizability. With respect to the embeddings learned for amino acids, the enrichment illustrated by Figure 3 highlights their capability to identify features not imparted by the BLOSUM matrix. At the terminal position in mhcRank’s top predicted peptides, all but one of the hydrophobic amino acids are enriched to comparable levels. Not only does this coincide with biological observations, but it also suggests that mhcRank has learned to identify, and favor, certain physiochemical properties of amino acids, such as hydrophobicity, despite no a priori knowledge [1]. Thus, even for the values of $k$ where the performance of mhcRank and netMHCpan were not significantly different, mhcRank is still superior as it is the more generalized predictor. Furthermore, the substantial improvement in performance over MHCflurry is aided by the implementation of CSSK. It is likely that these garner a boost in performance by identifying commonalities amongst protease cleave site motifs. Altogether, the proposed mhcRank demonstrates strong performance compared to existing methods, and could have vast applicability to aid drug and vaccine development.

References


Robert Leaman, Ph.D.\textsuperscript{1}, Qingyu Chen, Ph.D.\textsuperscript{1}, Alexis Allot, Ph.D.\textsuperscript{1}, Zhiyong Lu, Ph.D.\textsuperscript{1} 
\textsuperscript{1}NCBI/NLM/NIH, Bethesda, Maryland, USA

A substantial percentage of COVID-19 survivors experience debilitating symptoms long after the acute phase has resolved\textsuperscript{1}. This complex condition, called “long COVID” by patients, is not yet well understood, but it is becoming clear that many COVID-19 survivors will retain some degree of impairment for an extended period\textsuperscript{2}. Moreover, those at risk of long COVID include those not at high risk of mortality from COVID-19 due to mild/asymptomatic disease, youth, and good general health\textsuperscript{3}.

Despite the pressing need for information on long COVID, locating relevant research articles is challenging because the terminology used to refer to long COVID varies considerably. As a result, precise queries (e.g., “long COVID”) return limited results, but broader queries (e.g., “COVID sequelae”) suffer from low accuracy. We therefore created a comprehensive, searchable collection of long COVID articles, as an extension to LitCovid, a widely used literature hub with over 150,000 articles specific to COVID-19\textsuperscript{4}. We hope that the long COVID collection, updated weekly, will help researchers, healthcare professionals, and the general public keep up with the latest research.

Methods

We employed a human-in-the-loop process to identify all PubMed articles likely to be relevant to long COVID without requiring manual review of every article. Following previous definitions of long COVID\textsuperscript{1}, we defined an article to be relevant to long COVID if it considers “persistent symptoms and/or delayed or long-term complications [from COVID-19] beyond 4 weeks from the onset of symptoms.”

Starting from a small seed set of relevant articles, gathered manually, we split the data into four non-overlapping samples and trained a supervised classification model on each, using the LitSuggest web-based literature curation tool\textsuperscript{5}. We also extended our previous identifying variations in COVID-19 and SARS-CoV-2 terminology\textsuperscript{6} by creating a high-precision pattern-based named entity recognition (NER) system for mentions of long COVID.

We applied the LitSuggest classifiers and the NER system to all articles in the LitCovid portal\textsuperscript{3}, which includes all PubMed articles related to COVID-19. We then combined these relevance predictions and those of an external resource, CoronaCentral\textsuperscript{7}, into a single probabilistic prediction using triplet data programming\textsuperscript{8}, which estimates the accuracy of each input from the agreement rates between input pairs. Triplet data programming does not use labeled data or require training, enabling rapid analysis; our application generates predictions for all articles in LitCovid (154,429) in under 30 seconds.

We developed two powerful extensions to triplet data programming. First, the initial formulation only allows inputs to indicate relevant, irrelevant, or abstain; our method utilizes the full range of probabilities. Second, we improved the reliability of the accuracy estimates by ignoring input pairs where the 95% confidence interval on their agreement rate cannot rule out random chance.

Results

We applied our human-in-the-loop process to iteratively select articles for manual annotation – noting new long COVID terms and updating the NER patterns as needed. After 14 rounds, this iterative annotation resulted in 2,909 manually annotated articles (1,840 relevant, 1,069 irrelevant). Initial annotation rounds focused on articles that were likely to be relevant to long COVID; middle rounds focused on articles near the classification boundary and articles where inputs disagreed. The final rounds focused on likely false positives. In the final round, the f-score of the triplet data programming classifier is 0.831.

Of the 154,429 articles in LitCovid (as of July 26, 2021), we identified a total of 2,063 articles as relevant to long COVID. Figure 1a shows the number of articles on long COVID over time. Figure 1b compares the set of articles relevant to long COVID to the articles returned by querying PubMed for various long COVID terms and those in CoronaCentral. Note the low recall of precise queries (e.g., “post COVID syndrome”) and low precision of broad queries (e.g., post covid 19 sequelae), demonstrating the difficulty of locating long COVID articles by query alone. The NER system identified a total of 1,917 long COVID mentions in LitCovid (616 unique). Figure 1c shows the terms for long COVID found most frequently.
Figure 1. (a) Number of articles relevant to long COVID in PubMed per week. (b) Comparison of articles relevant to long COVID with articles returned by various PubMed queries and CoronaCentral. Precise queries (e.g., “post COVID syndrome”) return few results, broad queries (e.g., “post covid 19 sequelae”) return many irrelevant results. (c) Most commonly used terms for long COVID (case normalized and non-alphanumeric characters removed).

Conclusion


Acknowledgments

This research was supported by the Intramural Research Program of the National Library of Medicine at the NIH.

References

7. Lever J, Altman RB. Analyzing the vast coronavirus literature with CoronaCentral. PNAS. 2021 Jun 8;118(23).
VA’s National Telehealth Program: Contrasting Telehealth Modalities used by Primary Care and Mental Health Specialties during the COVID-19 Pandemic

Lucinda B. Leung, MD, PhD1,2; Lisa V. Rubenstein, MD, MSPH2,3,4; Erin Jaske MPH5; Chelle L. Wheat, PhD5; Karin M. Nelson, MD, MSHS5,6; Bradford L. Felker, MD5,7

1Center for the Study of Healthcare Innovation, Implementation, & Policy, VA Greater Los Angeles Healthcare System, Los Angeles, CA; 2Division of General Internal Medicine and Health Services Research, UCLA David Geffen School of Medicine, Los Angeles, CA; 3Department of Health Policy & Management, UCLA Fielding School of Public Health, Los Angeles, CA; 4RAND Corporation, Santa Monica, CA; 5VA Puget Sound Health Care System, Seattle, WA; 6Department of Medicine, University of Washington Medical School, Seattle, WA; 7Department of Psychiatry and Behavioral Sciences, University of Washington Medical School, Seattle, WA

Abstract
VA’s Clinical Resource Hub (CRH) program provided contingency staffing in primary care and mental health care using telehealth, with a focus on rural and/or under-resourced clinics. This retrospective cohort study examined 180,068 patients with at least one CRH visit between 10/01/2019 and 6/30/2021 (n=558,384 visits). We observed divergent use of CRH telehealth modalities between primary care and mental health patients over time. Results support growing telehealth capacity as a public health tool for national disasters.

Introduction
Telehealth is increasingly used to connect patients to providers and reduce access barriers related to distance or time.1 Veterans Health Administration (VHA)’s Clinical Resource Hub (CRH) telehealth program started October 2019 in all 18 VHA multi-state administrative regions.2 CRH aimed to address contingency staffing needs in primary care and mental health care through telehealth, with a focus on rural and/or under-resourced clinics. To obtain CRH support, clinics within a hub’s geographic and administrative catchment area requested contingency staffing; CRH hubs then could offer telehealth visits to patients at remote clinic clinics. Shortly after CRH implementation, the COVID-19 pandemic ensued, expanding the role of telehealth and providing an opportunity to observe a large telehealth program during a pandemic. This study tracks differential use of CRH visit types across specialties over time, comparing before and after the onset of the pandemic.

Methods
In this retrospective cohort study, we included all VHA patients with at least one CRH visit between 10/01/2019 and 6/30/2021 (n=558,384 visits). We counted CRH in-person, telephone, primary care clinic-based video, and patient home-based video visits3 by specialty (mental health, primary care, other). We compared primary care and mental health patterns of visit types used in the first year to focus on the early pandemic course. We used X2 and t-tests to determine significance using a two-tailed α of 0.05, with SAS version 9.4. This study was part of a VHA Office of Primary Care quality improvement project and exempt from Institutional Review Board review.

Results
Over 21 months, 180,068 patients (mean age 60.1[SD=16.14] years; 160,142[89%] men; 127,628[71%] White; 87,860[49%] rural) received CRH services. Only 895 (0.2%) visits were conducted in-person. Despite COVID-19, we observed a consistent, almost 3-fold increase in services. This was driven near-equally by telephone (254,177) and video (296,092) visits, the latter of which shifted largely from primary care clinic-based to patient home-based. Approximately half of visits were delivered by primary care (231,131) and mental health specialists (288,596).

With COVID-19’s onset, CRH primary care, compared to CRH mental health, relied increasingly on telephone (66,184[64%] versus video (37,543[36%]) visits, comprising typically of primary care clinic-based (30,879[30%]) rather than patient home-based (6,664[6%]) video. In contrast, CRH mental health increasingly relied on home-
based video visits (53,763[41%]), mostly replacing clinic-based video visits (30,879[30%]) while maintaining a steady, but lower rate of telephone contacts (40,081[31%]). (Figure 1)

Figure 1. Clinical Resource Hub Mental Health versus Primary Care Visits, 10/01/2019 and 9/30/2020

Conclusion

Early implementation findings suggest that unlike some healthcare programs, VHA’s CRH continued robust and increasing delivery of primary care and mental health care throughout COVID-19. Results support growing telehealth capacity as a public health tool for national disasters. CRH providers completed half of visits with rural patients, for whom disparities in telehealth use have been well documented. Geriatrics patients as old as 103 years, traditionally stereotyped against telehealth services, used CRH video visits.

CRH telehealth infrastructure is similar for VHA primary and mental health care; yet, observed differences in patterns of use (of telephone, primary care clinic-based, and patient home-based visits) between primary care and mental health specialties may reflect differences in care needs between patients and providers. For example, telephone visits may meet primary care needs for patient requests ranging from urgent problems (e.g., forms) to brief chronic condition care (e.g., medication refills). Video visits done from a primary care clinic may enable key elements of physical examination. Mental health visits, in contrast, may benefit from video-enabled visual cues and from patient comfort when communicating from home. Rather than focusing on baseline telehealth infrastructure as the main reason for differential telehealth utilization patterns, our data suggest the importance of considering intrinsic practice style differences and patient preferences when planning for telehealth expansion.

Acknowledgements

The authors would like to acknowledge Kendron Burnett and Matthew Rogers for their review of this abstract as senior leadership within the Clinical Resource Hubs.

References

Impact of Sex and Gender Disparities on Computational Phenotyping: A Potential Barrier to an Equitable Learning Health System

Rebecca T. Levinson, PhD¹, Jennifer R. Malinowski, PhD², Luke Rasmussen³, Suzette J. Bielinski, PhD⁴, Veronique L. Roger, MD, MPH⁵, Quinn S. Wells MD, PharmD⁶, Laura K. Wiley, PhD⁷
¹Heidelberg University Hospital, Heidelberg, Germany; ²Write InScite LLC, South Salem, NY; ³Northwestern University School of Medicine, Chicago, IL; ⁴Mayo Clinic, Rochester, MN; ⁵NHLBI, Washington, DC; ⁶Vanderbilt University Nashville, TN; ⁷University of Colorado School of Medicine, Aurora, CO

Introduction
Health disparities stem from both sex-based biological differences in disease prevalence, symptomology, and responses to treatment and gender-influenced differences in care access and delivery. The impacts of sex and gender on health and healthcare are well-documented, however little attention has been paid to how these disparities are propagated or exacerbated by cohort identification (i.e., “computational “phenotyping”) algorithms using electronic health record (EHR) data. Phenotyping algorithms can be as simple as a list of diagnosis codes, or involve more complex approaches (rule-based logic, NLP, machine learning, etc.) and multiple types of clinical data to identify the population of interest. Biases in the data used to construct these algorithms could result in the systematic underrepresentation of women, transgender, and non-binary patients.

Here we present an illustrative example of this effect in heart failure phenotyping. Heart failure (HF) is a complex syndrome with a heterogeneous presentation identified through history and clinical presentation rather than specific tests. However, HF subtype is determined by left ventricular ejection fraction (EF) measurement. Populations with preserved EF (HFpEF) skew female, older, and have more comorbidities than populations with reduced EF (HFrEF). EF is an easy and objective measure of HF, but when used in phenotyping algorithms it would preferentially identify HFrEF patients. This could lead to underidentification of women who are more likely to have HFpEF. Here we present a multimodal study of this effect through both EHR analyses and a systematic evidence review of the literature to determine the scale of this potential problem.

Methods
We applied three HF algorithms across the Vanderbilt University Medical Center Synthetic Derivative (de-identified EHR database): 1) the “eMERGE” algorithm; ² 2) a “Simple” algorithm with Boolean logic combining diagnosis codes, BNP lab tests, and loop diuretic prescriptions; and ³ a “Random Forest” algorithm using diagnoses, laboratory tests, medications, and text data. We identified the demographics and HF subtype for each population.

We also performed a systematic review of the heart failure phenotyping literature to identify the gender representation of published algorithms. ² Briefly, we identified studies from MEDLINE using terms for EHRs, phenotyping methodologies (e.g., NLP, machine learning, etc.), and heart failure. Blinded reviewers assessed studies in duplicate and extracted details about the algorithm and heart failure population demographics. Conflicts were adjudicated through discussion or the input of a third reviewer. We compared extracted demographics to the expected national female prevalence rate with Chi-square testing.

Results
EHR Analysis: A total of 54,608 patients were identified by one or more HF algorithms. All algorithms identified fewer women than expected (44% eMERGE, 45.1% Simple, 44% Random Forest) compared to national HF prevalence (51%) and representation in the Synthetic Derivative (53.7%). The Simple algorithm identified more women, possibly driven in part through increased identification of the HFpEF subtype (Figure 1).

Figure 1. HF phenotyping algorithms identify fewer women (pink) than expected from national prevalence estimates (black bar), and those that identify more HFpEF patients (green) also identify more women.
**Systematic Review:** Out of the 313 studies identified by our literature search, 25 studies developed 30 HF phenotyping algorithms. However, only 7 reported the demographics of their algorithm identified all-subtype HF population. Of these, 5 identified majority male populations and 6 identified cohorts that were statistically different from the expected national female prevalence rate of 51.6% by Chi-square testing (Figure 2). All 6 algorithms that were statistically different from the national prevalence rate used echocardiography data. The one algorithm that identified the expected patient proportion did not use echocardiography data.

**Discussion**

We found that both within a single institution and across the published literature the majority of HF algorithms identified significantly fewer women than expected from national prevalence estimates. We posit that one driver of this selection bias may be the reliance on echocardiography by the majority of algorithms that preferentially identifies HFrEF, a HF subtype more prevalent among men. Additional research is needed to confirm this effect and investigate other potential explanations for this observed bias.

Although we only examined the HF phenotype in this abstract, we expect that sex and gender disparities will affect algorithms across many phenotypes. Diagnosis codes are the most common data element used for phenotyping but rely on patients receiving a proper diagnosis. Research has shown that women are often subject to long diagnosis delays, reducing the number of women identified and potentially selecting for women with more severe or later disease presentations. Algorithms also use medications and procedures as a way to infer case status, but in the case of acute myocardial infarction there has been underuse of evidence-based treatment in women that would reduce the efficacy of these data to identify women. There are similar sex- and gender-based differences in laboratory and clinical measurements used in phenotyping algorithms, which are often more pronounced in transgender patients.

The potential impact of sex and gender bias is significant, affecting not only research participation (and therefore study finding generalizability), but also clinical care as similar approaches are used throughout the learning healthcare system including clinical decision support, population health management, and clinical quality evaluation. In order to minimize the effect of potential bias within phenotyping algorithms, we must: 1) understand how sex and gender have been considered in existing algorithms; 2) develop a framework for evaluating the impact of health disparities on phenotyping; 3) develop standards for reporting algorithms that allow the proactive assessment of potential bias.

**References**

Digital Readiness: Are Low-Income Populations Prepared for Telehealth?

Sunny C. Lin, PhD, MS,1 Julie Reeder, PhD, MPH, CHES2, Mary Vest, BA1, Kathy Harris, PhD3, Jill Castek, PhD, MS,4
1OHSU-PSU School of Public Health, Portland, OR; 2Oregon WIC, Portland, OR; 3Portland State University, Portland, OR; 4University of Arizona, Tucson, AZ

Introduction
Videoconferencing is increasingly viewed as a way to make health services accessible and affordable. Yet videoconferencing may not benefit their intended populations equitably if disparities exist in access to Internet and videoconference devices (i.e., smartphones, computers with webcams) and/or preferences for videoconferencing. Despite the widespread belief that videoconferencing will improve the accessibility of health services, few studies have examined whether people in marginalized communities have the interest or ability to use videoconferencing for health services. While prior studies have found that low-income populations are increasingly interested in using technology to access medical information asynchronously such as through patient portals, little is known about whether videoconferencing is a viable method for expanding synchronous health services access, and what help people need when using videoconferencing for health services access (1). Understanding this knowledge gap is critical to crafting telehealth policies and interventions that do not exacerbate existing disparities (2). The Oregon’s Women, Infant, and Children’s Supplemental Nutrition Program (WIC) offers supplemental nutrition, education, breastfeeding, and lactation consultant services to low-income communities in Oregon. We conducted an assessment of participants’ Internet and device access, experience with, needs for supporting, and preference for using videoconferencing compared to telephone and in-person visits for WIC services.

Methods
All currently-enrolled English, Spanish, or Russian-speaking Oregon WIC participants were sent a text message inviting them to participate in the survey in early December 2020. The survey remained open from December 2020 through March 2021. The survey was offered in three different languages: Spanish, Russian, and English. To improve participation rates among participants who may not have access to Internet-enabled smart phones or computers, participants with an education level of less than high school completion were also given the option to complete the survey over the phone. Due to resource constraints, we limited the phone survey option to those with education level of less than high school completion only. To determine participants’ Internet access, we asked whether they had the following kinds of Internet access where they live: monthly internet service plan (ISP), monthly cell phone data plan, or prepaid cell phone plan. To determine participants’ device access, we asked what devices participants had and how often they used each (i.e., don’t have, use almost every day, monthly, rarely). To determine participants’ preferences for videoconference modality, we asked participants, “If video conferencing were to become a regular option for WIC appointments in the future, would you choose to: go back to in-person appointments, keep having telephone calls, or have a videoconference appointment using a smartphone, computer, or tablet.” Chi-squared tests were used to determine answers to these questions differed by characteristics (i.e., language, urban/rural location, race, and ethnicity). Additional survey questions asked whether participants ever ran out of mobile data and how often, accessed data outside their home due to unreliable Internet access, prior experience with videoconferencing for telehealth, and what help participants need to access videoconferencing.

Results
Out of 62,418 currently-enrolled WIC participants, 61,248 WIC participants were identified from current records as those who spoke English (n=51705, 82%), Spanish (n=9033, 14%), or Russian (n=502, 1%). Of those, 44,927 were sent a text message invitation to participate in the survey 76% in English (n=5713), 23% in Spanish (n=1740), and 1% in Russian (n=103), 1,704 Spanish (n=9033), or Russian (n=502). Fifty-six English speakers, 5 Russian speakers, and 238 Spanish speakers opted to take the survey by telephone. Survey participants were similar to the overall WIC population by geography and race/ethnicity. Survey responses about Internet and device access, and preferred modality are presented for all respondents and by language, geography, race, and ethnicity in Table 1. In addition, we found that 18% of participants report running out of mobile data every month, 14% every few months, 12% once or twice a year, and 56% never. Nearly half (48%) of participants said they sometimes need to connect to the Internet in creative places outside their homes such as at a school, public library, church, or the parking lot of a business with a strong wireless Internet signal. When participants were asked about their
experiences with and needs to support telehealth, 47% indicated they had ever used telehealth and liked it, 37% indicated they had ever used telehealth but did not like it, 10% reported never having used telehealth but would try it, and 6% reported never having used telehealth and were not interested in using it. When asked what help they would need for accessing video visits for WIC services, the most commonly reported needs were: finding a private or quiet enough place (40%), reliable Internet access at home (17%), support logging in to the visit (12%), device access (11%), and support downloading the video conference platform or app (10%). Of those that responded to the survey in Russian or English, 12% reported needing help navigating the appointment in English.

Table 2. Survey Responses by Respondent Characteristics (n=7556)

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>N</th>
<th>% Of Subgroup with Internet Access</th>
<th>% of Subgroup Who Use Device Daily</th>
<th>% of Subgroup with Preferred Modality</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Have Monthly ISP</td>
<td>Have Monthly Mobile</td>
<td>Have Prepaid Mobile</td>
</tr>
<tr>
<td>All Respondents</td>
<td>7,556</td>
<td>81%</td>
<td>74%</td>
<td>19%</td>
</tr>
<tr>
<td>Language</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>English</td>
<td>5,713</td>
<td>84%***</td>
<td>80%***</td>
<td>16%***</td>
</tr>
<tr>
<td>Spanish</td>
<td>1,740</td>
<td>73%***</td>
<td>55%***</td>
<td>27%***</td>
</tr>
<tr>
<td>Russian</td>
<td>103</td>
<td>64%***</td>
<td>72%***</td>
<td>7%***</td>
</tr>
<tr>
<td>Geography</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Urban</td>
<td>3,580</td>
<td>83%***</td>
<td>74%***</td>
<td>17%***</td>
</tr>
<tr>
<td>Suburban</td>
<td>2,639</td>
<td>81%***</td>
<td>75%***</td>
<td>18%***</td>
</tr>
<tr>
<td>Rural</td>
<td>1,284</td>
<td>78%***</td>
<td>74%***</td>
<td>25%***</td>
</tr>
<tr>
<td>Missing</td>
<td>53</td>
<td>53%***</td>
<td>54%***</td>
<td>36%***</td>
</tr>
<tr>
<td>Race</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>5,308</td>
<td>83%***</td>
<td>78%***</td>
<td>17%***</td>
</tr>
<tr>
<td>Not White</td>
<td>2,248</td>
<td>76%***</td>
<td>65%***</td>
<td>23%***</td>
</tr>
<tr>
<td>Ethnicity</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hispanic</td>
<td>3,180</td>
<td>77%***</td>
<td>66%***</td>
<td>22%***</td>
</tr>
<tr>
<td>Not Hispanic</td>
<td>4,309</td>
<td>85%***</td>
<td>80%***</td>
<td>17%***</td>
</tr>
<tr>
<td>Missing</td>
<td>67</td>
<td>72%***</td>
<td>72%***</td>
<td>21%***</td>
</tr>
</tbody>
</table>

Notes: ISP = Internet Service Provider; Chi-square test *** p<0.001, **p<0.01, *p<0.05 1Percent denominator includes 329 missing data on monthly ISP; 2Percent denominator includes 780 missing data on monthly mobile data plan; 3Percent denominator includes 1,297 missing data on monthly prepaid mobile data plan; 4Percent denominator includes 246 missing data on smart phone use; 5Percent denominator includes 1000 missing data on tablet use; 6Percent denominator includes 841 missing data on computer use; 7Percent denominator includes 16 missing data on preferred modality

Discussion

In this study, we found significant differences in videoconference access and preference by language, geography, race, and ethnicity, suggesting moving too quickly to advanced technologies (e.g., video versus phone visits) for health services related may alienate a large portion of low-income populations. Our findings suggest that possible patient-centered videoconference solutions such as designated quiet spaces, childcare, device and hotspot libraries, not requiring visual components to synchronous appointments, and language interpreters. Limitations include potential bias due to survey non-response, and limited generalizability outside of the Oregon area. Future research ought to examine association between videoconferencing preference and COVID-19 prevalence and perceived threat as well as correlation and prediction with outcomes such as satisfaction and adherence, or type of care received.

References

COVID-19 Mortality Prediction among Patients with Cancer Using a Large National Cohort

Feifan Liu, PhD1, Noha Sharafeldin, MD, MSc, PhD2, Vithal Madhira, MS3, Kate Bradwell, PhD4, Qianqian Song, PhD5, Benjamin Bates, MD6, Yu Raymond Shao, MD, PhD7, Jing Su, PhD8, Alfred Jerrod Anzalone, MS9, Tim Bergquist, PhD10, Sarah L. Cutrana, MD, MPH1, Ben S. Gerber, MD, MPH1, Peter N. Robinson, MD, PhD11, Justin Guinney, PhD10, Umit Topaloglu, PhD, FAMIA5
1University of Massachusetts Medical School, Worcester, MA; 2University of Alabama at Birmingham, Birmingham, AL; 3Palilla Software, Reno, NV; 4Palantir Technologies, Denver, CO; 5Wake Forest School of Medicine, Winston Salem, NC; 6Rutgers Center for Pharmacoepidemiology and Treatment Science, New Brunswick, NJ; 7Duke Cancer Institute, Durham, NC; 8Indiana University School of Medicine, Indianapolis, IN; 9University of Nebraska Medical Center, Omaha, NE; 10Sage Bionetworks, Seattle, WA; 11The Jackson Laboratory, Farmington, CT

Introduction
Globally, more than 2 million people have died due to COVID-19 infection.1 An international study reported an increased COVID-19-associated mortality rate among cancer patients compared to non-cancer patients (20% vs. 11%, P = .006).2 Individualized risk prediction for adverse COVID-19 outcomes (all-cause mortality or discharge to hospice) among patients with cancer will help inform evidence-based clinical management, and this is of even more heightened importance amongst patients undergoing active cancer treatment.3 Risk stratification based on clinical and demographic factors can also facilitate more efficient utilization of limited resources in a pandemic setting through proper patient triaging.4 Although COVID-19 mortality prediction models have been published for the general population, patients with cancer remain understudied and are limited to a small sample size of under 1,000.4,5 In this study, we analyzed a large national cohort of cancer patients and developed a COVID-19 mortality prediction model using classic machine learning algorithms.

Methods
We constructed our study cohort from 3.8 million patients included in the National COVID Cohort Collaborative (N3C) electronic health record (EHR) data repository. As of February 25, 2021, 45,511 patients had a diagnosis of cancer preceding a COVID positive status. We excluded patients without records of medication or lab tests at the first COVID positive date. Our final cohort consisted of 8,336 patients. We extracted 91 predictors including demographics (8), medications (46), Charlson Comorbidity Index7 (1), comorbidities (18), lab test results (18) at the time of first COVID positive diagnosis (allowing for early intervention), as well as mortality outcomes (death or discharge to hospice). Choices of predictor variables were based on clinical oncologist expert opinion as well as previous studies on mortality prediction for general population8 and patients with cancer4,5. All the N3C data were transformed into a harmonized Observational Medical Outcomes Partnership (OMOP) common data model, and our cohort and concept definitions are based on the Observational Health Data Sciences and Informatics (OHDSI) standard concept sets. Patients were randomly selected for a 70-30 training-test split. We trained predictive models using two commonly used classic machine learning algorithms, random forest (RF) and logistic regression (LR), because they were easily interpretable. The GridSearch strategy with 10-fold cross validation was used for hyperparameter optimization. To address the data imbalance issue, cost-sensitive9 approaches, which put more penalties on prediction errors from the minority class (patients with mortality outcomes), were explored during model training. Permutation importance scores were calculated on the test data to assess predictors’ contribution to model decisions.

Results
Among our cohort of 8,336 COVID patients with cancer, 18% (n=1503) either died or were

Figure 1. Receiver operating characteristic (ROC) curve (left) and precision recall curve (right) for RF and LR on the test data.
discharged to hospice. The RF model outperformed LR model for mortality prediction, yielding an area under the receiver operating characteristic curve (AUROC) of 0.847 (95% confidence interval [CI]: 0.826-0.867) vs. 0.833 (95% CI: 0.810-0.854) for LR. The area under the precision recall curve (PR_AUC) was 0.613 (95% CI: 0.563-0.663) for RF vs. 0.574 (95% CI: 0.521-0.626) for LR. ROC and PR curves for both models are shown in Figure 1. RF obtained a better precision of 74.4% (95% CI: 68.3%-79.8%) vs. 46.3% (95% CI: 41.9%-50.4%) for LR, while LR yielded a better recall of 60.8% (95% CI: 56%-65.6%) vs. 40.6% (95% CI: 35.6%-45.3%) for RF. The performance differences between two models on the aforementioned four metrics were all statistically significant based on the Student's t-test (p<0.05). Among the top 30 predictors identified by each model, 23 were shared between the two models, which included components of complete blood counts (hemoglobin, platelets, white blood cells, lymphocytes, neutrophils), labs related to nutritional status and hepatic/renal function (albumin, aspartate aminotransferase, bilirubin, blood urea nitrogen), chemistry labs (sodium, potassium, glucose), steroids, antiarrhythmic medication (amiodarone), antiviral medication (remdesivir), pressors (norepinephrine, vasopressin), anti-hypertensive medications (angiotensin converting enzyme inhibitor, angiotensin receptor blocker), age, Charlson Comorbidity Index, and some individual comorbidities (dementia, paralysis, metastatic solid tumor). LR model also identified “not_Hispanic_or_Latino” as one of the top 30 predictors. Figure 2 shows the top 10 predictors identified by RF and LR respectively.

**Discussion**

We demonstrated that EHR data from a large national cohort can be leveraged for predicting severe COVID outcomes (death or discharge to hospice within one year from the first COVID positive diagnosis) in patients with cancer. We found that random forest performed slightly better than logistic regression on this task, but the latter achieved a better sensitivity. It is possible that ensemble methods combining different machine learning models can further improve prediction performance.

Complete blood count parameters, albumin and age were among the top 10 shared predictors using both models highlighting the utility of commonly used, inexpensive lab tests in predicting overall mortality in patients with COVID-19 and cancer. Lab tests related to liver/kidney functions are also important predictors. The current study did not take into consideration longitudinal features and only included a limited set of commonly used OMOP concepts for feature extraction. Our study focused on one-year mortality - further studies with other lengths of follow-up might also be informative. Future studies will address the above limitations, assess potential model biases, and explore additional machine learning algorithms including deep neural networks.

**References**

Leveraging Transfer Learning to Analyze Opinions, Attitudes, and Behavioral Intentions Toward COVID-19 Vaccines

Sirus Liu, PhD1, Jialin Liu, MD2
1Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, Tennessee; 2Department of Medical Informatics, West China Hospital, Chengdu, China

Introduction
As a significant global health threat, long-term control of COVID-19 relies on the development and acceptance of a preventive vaccine.1,2 However, according to a recent survey, only 51% of the 10,093 adults in the United States indicated that they would be willing to receive the COVID-19 vaccine when it becomes available3, hardly achieving the recommended threshold of 70% to reach herd immunity.4 While previous studies have explored knowledge in other vaccines using deep learning methods,5 questions related to COVID-19 vaccines remain unknown: What is the prevalence of opinions on the social media platform? How many tweets express attitudes or behavioral intention to take vaccines? Which topics are mostly associated with these contents? To answer these questions, we developed transfer learning models to detect content expressed user opinions, attitudes and behavioral intentions toward COVID-19 vaccines. We then conducted a temporal analysis to explore trends and developed probabilistic topic models to obtain the most important topics. We believe this study will be of great benefit to the timely rollout of the vaccines by extracting the latest public opinions, attitudes, and behavioral intentions to tailor promotion programs.

Methods
We used a combination of keywords and hashtags related to COVID-19 vaccines: “(#covid OR covid OR #covid19 OR covid19) AND (#vaccine OR vaccine OR #vaccine OR vaccine OR vaccinate OR immunization OR immune OR vax)” to collect tweets in English published from November 1, 2020 to January 31, 2021. We randomly selected 5,000 tweets, annotated by two independent reviewers (SL and JL). For each tweet, we labeled whether it included an opinion (yes/no), attitude (positive/negative/unknown), or a behavioral intention (yes/no/unknown).

After the data preprocessing, we split the annotated dataset into three parts: training (60%), validation (20%), and testing (20%). We applied traditional machine learning algorithms (logistic regression, random forest, and support vector machine) with Text Frequency Inverse Document Frequency (TF-IDF). We used the BERT-base-cased as the pretrained language model, the “BERT for sequence classification” model as the pretrained classification model, and the Adam algorithm with weight decay (AdamW) as the optimizer. The BERT models were generated using the huggingface package in Python 3 and developed with the Google Colab platform with a high-RAM GPU. We reported outcomes with 1,000 rounds of bootstrapping. The primary outcome was macro F1. We performed the Nemenyi test to compare the F1 values.

We applied the optimal models to predict the unlabeled data. We calculated the daily proportion of tweets classified as containing opinions to the total number of tweets and the percentage of the tweets predicted to exhibit a particular attitude or behavioral intention to all tweets indicating attitudes or behavioral intentions. To test the variability of this ratio over time, we performed the Augmented Dickey-Fuller (ADF) test with a significant threshold of \(p<0.05\). We conducted the latent Dirichlet allocation (LDA) analysis to extract the main topics.

Results
We annotated 5,000 tweets from 4,796 unique users (\(\kappa=0.76\)). The prediction performances of models are presented in Table 1. The transfer learning model significantly outperformed the machine learning models (\(p<0.05\)), which achieved F1 values of 0.792, 0.578, and 0.614 for three tasks.

We collected 2,678,372 tweets related to COVID-19 vaccines posted by 841,978 unique users. The prevalence of tweets expressing opinions was 0.222 [0.202, 0.245] and was significantly stable over time (ADF=-4.341, \(p<0.001\)). The rates of negative attitudes and positive attitudes were 0.754 [0.707, 0.795] and 0.246 [0.204, 0.293], respectively, and varied over time (ADF=-1.137, \(p=0.700\)). The rates of tweets indicating that users would not vaccinate and would vaccinate were 0.342 [0.229, 0.461] and 0.652 [0.539, 0.771], respectively, and also varied over time (ADF=-0.980, \(p=0.761\)). In addition, we observed a substantial increase in the prevalence of tweets expressing positive behavioral intention starting from mid-December 2020.
We extracted ten topics for each category. For negative attitudes, dominant topics included the concerns about the safety issues for children and pregnant women; unknown side effects; rushing the development process; the virus mutation; and trust in the government or scientists. Positive attitudes came primarily from news of effective trial results and users hoped that COVID-19 vaccines could end the pandemic as well as the desire to return to a normal life. Negative behavioral intentions came from the concerns of long-term and unknown side effects; non-acceptance of forced vaccination; and the lack of need for vaccination against a disease with a low lethality rate. For positive behavioral intentions, the dominant topic was that users would like to become immune to the virus and stay healthy.

**Discussion**

While previous studies have classified sentiment in COVID-19 related tweets, in our study, the task of classifying attitudes towards COVID-19 vaccines is more difficult than the common sentimental analysis. We not only focused on the sentiment of tweets, but also considered simultaneously whether the object of the sentiment was the COVID-19 vaccine. During annotation, we noticed that some tweets contained positive words used to describe what would happen after the vaccine rollout, but also stated negative attitudes toward the vaccine itself, such as lack of trust.

This study has several limitations. First, users of the Twitter platform are not representative of the entire public. The Twitter platform is usually considered to gather more anti-vaccinators and spread misinformation. Understanding their perceptions is a necessary step to tailor vaccine promotion education materials, to have a better chance of effectively changing their behavior. Second, some terms extracted from the topic modeling were difficult to infer. Given the complex situation of behavioral intentions toward the COVID-19 vaccines, our next step is to conduct a theory-based content analysis to understand psychological aspects of why some people will not get the vaccines.

In this study, we presented an annotated corpus of 5,000 tweets and demonstrated the capability to use transfer learning to identify public opinions, behavioral intentions, and attitudes towards COVID-19 vaccines from social media. In addition, we explored temporal trends in attitudes and behavioral intentions on a larger dataset with 2,678,372 tweets from 2020.11.01-2021.01.31. We found that the LDA technique is useful to extract topics. Overall, we provided a method to analyze the perceptions of COVID-19 vaccines from the real-time data, which could be used to tailor the education programs and other interventions to promote COVID-19 vaccine acceptance urgently.

**Acknowledgements**

The authors do not have conflicts of interest related to this study.

**References**


---

Table 1. Metrics (mean [95%CI]) of outcomes in classifying tweets related to covid-19 vaccines.

<table>
<thead>
<tr>
<th>Task</th>
<th>Model</th>
<th>Recall</th>
<th>Precision</th>
<th>F1</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Opinions</strong></td>
<td>BERT</td>
<td>0.762 [0.759, 0.766]</td>
<td>0.862 [0.858, 0.866]</td>
<td><strong>0.792</strong> [0.789, 0.795]</td>
<td>0.854 [0.852, 0.856]</td>
</tr>
<tr>
<td></td>
<td>LR</td>
<td>0.774 [0.770,0.779]</td>
<td>0.757 [0.753, 0.762]</td>
<td>0.764 [0.761, 0.767]</td>
<td>0.807 [0.805, 0.810]</td>
</tr>
<tr>
<td></td>
<td>RF</td>
<td>0.754 [0.750, 0.758]</td>
<td>0.732 [0.728, 0.735]</td>
<td>0.740 [0.737, 0.743]</td>
<td>0.783 [0.781, 0.786]</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>0.767 [0.764, 0.771]</td>
<td>0.752 [0.748, 0.755]</td>
<td>0.758 [0.755, 0.761]</td>
<td>0.803 [0.801, 0.806]</td>
</tr>
<tr>
<td><strong>Attitudes</strong></td>
<td>BERT</td>
<td>0.529 [0.521, 0.536]</td>
<td>0.698 [0.686, 0.710]</td>
<td>0.578** [0.572, 0.584]</td>
<td>0.873 [0.871, 0.875]</td>
</tr>
<tr>
<td></td>
<td>LR</td>
<td>0.475 [0.468, 0.482]</td>
<td>0.530 [0.520, 0.541]</td>
<td>0.495 [0.490, 0.500]</td>
<td>0.859 [0.856, 0.861]</td>
</tr>
<tr>
<td></td>
<td>RF</td>
<td>0.518 [0.511, 0.526]</td>
<td>0.558 [0.545, 0.570]</td>
<td>0.508 [0.502, 0.514]</td>
<td>0.830 [0.827, 0.833]</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>0.506 [0.498, 0.514]</td>
<td>0.551 [0.541, 0.562]</td>
<td>0.523 [0.517, 0.530]</td>
<td>0.863 [0.860, 0.865]</td>
</tr>
<tr>
<td><strong>Behavioral Intentions</strong></td>
<td>BERT</td>
<td>0.562 [0.549, 0.575]</td>
<td>0.734 [0.716, 0.752]</td>
<td><strong>0.614</strong> [0.606, 0.622]</td>
<td>0.961 [0.960, 0.962]</td>
</tr>
<tr>
<td></td>
<td>LR</td>
<td>0.472 [0.461, 0.483]</td>
<td>0.725 [0.699, 0.752]</td>
<td>0.527 [0.519, 0.536]</td>
<td>0.801 [0.800, 0.802]</td>
</tr>
<tr>
<td></td>
<td>RF</td>
<td>0.447 [0.437, 0.457]</td>
<td>0.577 [0.543, 0.611]</td>
<td>0.466 [0.457, 0.476]</td>
<td>0.935 [0.934, 0.937]</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>0.469 [0.458, 0.479]</td>
<td>0.710 [0.684, 0.737]</td>
<td>0.523 [0.513, 0.533]</td>
<td>0.950 [0.948, 0.951]</td>
</tr>
</tbody>
</table>

Abbreviations: **p<0.05 in the Nemenyi test; BERT: Bidirectional Encoder Representations from Transformers; LR: Logistic Regression; RF: Random Forest; SVM: Support Vector Machine.**
A Longitudinal Study of Burnout and Clinical Workload Measured With Electronic Health Record Audit Logs

Sunny S. Lou MD PhD, Daphne Lew PhD MPH, Derek R. Harford BA, Chenyang Lu PhD, Bradley A. Evanoff MD MPH, Jennifer G. Duncan MD, Thomas Kannampallil PhD

Washington University School of Medicine, St Louis, MO.

Introduction: Burnout is primarily a work-related phenomenon that affects nearly 50% of physicians across the United States, with consequences for the health of both physicians and their patients (1). The causal contributors to this high rate of burnout and potential mitigating steps are not well understood (1). Previous cross-sectional surveys paired with electronic health record (EHR) measurements have shown that increased inbox message use and afterhours work are associated with burnout among outpatient primary care physicians (2–4). However, the workload-related factors that contribute to burnout among inpatient clinicians, who make up a large fraction of the workforce, are less explored. Furthermore, the longitudinal evolution of burnout with varying workloads has not been characterized. Resident physicians are an ideal population to explore both of these issues; residents rotate between inpatient and outpatient settings and experience large shifts in workload and workflow that can be thought of as natural experiments. The goal of this study was to measure resident workload across varying clinical rotations using EHR audit log data and investigate whether specific types of workload are more associated with burnout.

Methods: We conducted a six-month prospective longitudinal observational study of intern physicians from Medicine, Pediatrics and Anesthesiology at a large academic medical center. These interns rotated between inpatient medicine, intensive care, ambulatory, and elective months. Enrollment occurred during 9/2020-10/2020. Consented participants were asked to complete monthly burnout surveys at the end of each rotation for six months. Burnout was measured using the Stanford Professional Fulfillment Index (PFI); this instrument was chosen because it is validated to assess burnout over short time scales (5). PFI scores range from 0 to 4, and scores > 1.33 were considered to have met criteria for burnout.

Workload was measured from the participants’ EHR audit logs (Epic Systems, Verona, WI). Data for the month prior to survey completion was associated with each survey response. EHR audit logs record timestamps and patient identifiers each time a clinician views or modifies elements of the patient record. Although designed to monitor security access, they are also a fingerprint of a clinician’s work habits and have been broadly adopted by the informatics community to measure clinician workload and behavior (6). We measured time spent on each EHR-based action as the time interval between that action and the subsequent action, ignoring time differences greater than five minutes which were assumed to represent inactivity (7). Actions were categorized to orders, notes, and data review by a physician member of the study team. We computed the following monthly workload metrics: total EHR time, total afterhours EHR time (time spent between 6pm and 6am), and total time spent on inbox messages. We also computed the following normalized monthly metrics: average time spent reviewing data per patient per day, average time spent on notes per patient per day, average number of orders placed per patient per day, and average number of patients per day for whom the participant placed either a note or order, used as a proxy for patient load.

A linear mixed effect model was used to assess the relationship between workload metrics and the participants’ monthly burnout scores. Workload metrics were included as fixed effects with a random intercept for each participant. Specialty was included as a fixed effect to adjust for potential confounding. All workload metrics were initially included in the model and stepwise model selection was used. P-values were computed with Satterthwaite’s method. Workload measurements were made with custom python code. Analysis was conducted in Python and R.

Results: 75 of 104 eligible participants elected to enroll in the study (72.1%). The completion rate for subsequent surveys was 76.0% across all participants. Demographic characteristics and average EHR workload metrics for the study population are shown in Table 1. 42.7% of the participants were burned out upon enrollment into the study. Of the participants who completed at least two surveys, 42.0% (29/69) experienced at least one transition between burned out and not burned out or vice versa, and 24.6% (17 / 69) demonstrated at least two transitions.

Results from the multivariable linear mixed effect model are shown in Table 2. Total EHR time, average number of patients seen per day, and average time spent reviewing data per patient per day were significantly associated with burnout. For example, after adjusting for other factors, every additional hour per patient per day spent reviewing data was associated with an increase in burnout score by 0.48 (95% CI 0.08 – 0.89) points. Afterhours EHR time, average number of orders per patient per day, and average time spent on notes per patient per day were not significantly associated with burnout. No statistically significant difference in burnout between specialties was observed.
**Discussion:** We conducted a prospective longitudinal observational study of the time course of burnout among interns engaging in both inpatient and outpatient medicine, and measured the association between their burnout and objective measures of workload as captured by EHR audit logs. We demonstrate that burnout is not static and often varies over time, likely as a function of recent workload and related personal factors. We found that increased patient load, increased EHR time, and increased time spent reviewing EHR data per patient are associated with increased burnout. It is possible that these variables are proxies for other contributors such as total work hours or patient complexity. In contrast to previous findings among primary care physicians (2–4), inbox message volume and afterhours time were not associated with burnout.

This study is one of the first to associate objective EHR-based measures of workload to burnout among clinicians working across both inpatient and outpatient settings, and also is one of the first longitudinal studies of burnout. Our findings might help inform EHR usability and design, and suggest that interventions targeted to improve the efficiency of data review might be highest yield to reduce EHR-related burnout. In addition, workplace interventions to reduce patient load or work hours may be worth considering where feasible. This study was limited to a single institution with incomplete participation that could contribute to response bias, although our participation rate was high. Also, we were unable to measure non-EHR work and our EHR workload measures were imperfect; nonetheless, they likely correlated with true workload.

<table>
<thead>
<tr>
<th>Specialty – Medicine</th>
<th>n (%) for categorical / median (IQR) for numeric</th>
<th>Coefficient (95% CI)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Specialty: Medicine</td>
<td>Reference</td>
<td>-0.246 (-0.598 - 0.106)</td>
<td>0.276</td>
</tr>
<tr>
<td>Pediatrics</td>
<td>-0.246 (-0.598 - 0.106)</td>
<td>0.276</td>
<td></td>
</tr>
<tr>
<td>Anesthesiology</td>
<td>-0.032 (-0.343 - 0.407)</td>
<td>0.119</td>
<td></td>
</tr>
<tr>
<td>Total EHR time / mo</td>
<td>0.002 (0.000 – 0.003)</td>
<td>0.024</td>
<td></td>
</tr>
<tr>
<td>Avg patient load / day</td>
<td>0.059 (0.024 – 0.094)</td>
<td>&lt;0.001</td>
<td></td>
</tr>
<tr>
<td>Avg review time / pt / day</td>
<td>0.480 (0.077 – 0.882)</td>
<td>0.020</td>
<td></td>
</tr>
<tr>
<td>Specialty: Medicine</td>
<td>Reference</td>
<td>-0.246 (-0.598 - 0.106)</td>
<td>0.276</td>
</tr>
<tr>
<td>Pediatrics</td>
<td>-0.246 (-0.598 - 0.106)</td>
<td>0.276</td>
<td></td>
</tr>
<tr>
<td>Anesthesiology</td>
<td>-0.032 (-0.343 - 0.407)</td>
<td>0.119</td>
<td></td>
</tr>
<tr>
<td>Total EHR time / mo</td>
<td>0.002 (0.000 – 0.003)</td>
<td>0.024</td>
<td></td>
</tr>
<tr>
<td>Avg patient load / day</td>
<td>0.059 (0.024 – 0.094)</td>
<td>&lt;0.001</td>
<td></td>
</tr>
<tr>
<td>Avg review time / pt / day</td>
<td>0.480 (0.077 – 0.882)</td>
<td>0.020</td>
<td></td>
</tr>
<tr>
<td>Specialty: Medicine</td>
<td>Reference</td>
<td>-0.246 (-0.598 - 0.106)</td>
<td>0.276</td>
</tr>
<tr>
<td>Pediatrics</td>
<td>-0.246 (-0.598 - 0.106)</td>
<td>0.276</td>
<td></td>
</tr>
<tr>
<td>Anesthesiology</td>
<td>-0.032 (-0.343 - 0.407)</td>
<td>0.119</td>
<td></td>
</tr>
<tr>
<td>Total EHR time / mo</td>
<td>0.002 (0.000 – 0.003)</td>
<td>0.024</td>
<td></td>
</tr>
<tr>
<td>Avg patient load / day</td>
<td>0.059 (0.024 – 0.094)</td>
<td>&lt;0.001</td>
<td></td>
</tr>
<tr>
<td>Avg review time / pt / day</td>
<td>0.480 (0.077 – 0.882)</td>
<td>0.020</td>
<td></td>
</tr>
<tr>
<td>Specialty: Medicine</td>
<td>Reference</td>
<td>-0.246 (-0.598 - 0.106)</td>
<td>0.276</td>
</tr>
<tr>
<td>Pediatrics</td>
<td>-0.246 (-0.598 - 0.106)</td>
<td>0.276</td>
<td></td>
</tr>
<tr>
<td>Anesthesiology</td>
<td>-0.032 (-0.343 - 0.407)</td>
<td>0.119</td>
<td></td>
</tr>
<tr>
<td>Total EHR time / mo</td>
<td>0.002 (0.000 – 0.003)</td>
<td>0.024</td>
<td></td>
</tr>
<tr>
<td>Avg patient load / day</td>
<td>0.059 (0.024 – 0.094)</td>
<td>&lt;0.001</td>
<td></td>
</tr>
<tr>
<td>Avg review time / pt / day</td>
<td>0.480 (0.077 – 0.882)</td>
<td>0.020</td>
<td></td>
</tr>
</tbody>
</table>

Table 1 – Characteristics of the study population.

Table 2 – Estimated beta coefficients, 95% confidence intervals, and p-values for the fixed effects in the linear mixed effects model for monthly PFI burnout score.

**References:**


Feasibility of using machine learning on insurance claims to identify correlates of lower extremity amputation in insured adults with type 2 diabetes who initiate treatment with sodium-glucose co-transporter 2 (SGLT-2) inhibitors

Yuan Luo1, Andrew Cooper1, Raymond Kang1, Ronald T Ackermann1
1Feinberg School of Medicine, Northwestern University, Chicago, IL, US

Introduction

Patients with type 2 diabetes (T2D) are at increased risk for foot ulcer and amputation. Risk is increased in the setting of poorly controlled blood pressure, peripheral arterial disease (PAD), or distal polyneuropathies that limit protective sensation and predispose to low tissue perfusion, foot injury, and infection. In 2016, FDA issued a safety alert regarding amputations associated with canagliflozin, one of 4 currently FDA-approved diabetes medications in the class known as sodium-glucose co-transporter 2 (SGLT-2) inhibitors. The alert was based on a 2-fold increase in risk observed over 2.4 years of median follow up in the Canagliflozin Cardiovascular Assessment Study (CANVAS).

Results from trials of SGLT2 agents have observed elevated amputation risk with unknown mechanism, and it is possible that other patient characteristics or exposures explain the incremental risk observed in CANVAS1. The purpose of our study is to identify both known and new risk factor combinations from insurance claims data that can be integrated in a single model to predict excess amputation risk in individuals being considered for initiation of a SGLT-2 for their diabetes. Note that the focus of this study is on models that are easy to interpret and features broadly applicable and obtainable from an administrative database.

Methods

Study Population: The SGLT-2 new user cohort included all adults (age 18+), who had evidence of type 2 diabetes (T2D), had a first pharmacy fill for a SGLT-2 between 1/1/2013 and 6/30/2018, and covered by a large health insurer for at least 6 months before and 6 months after the first fill date (index date). Figure 1 presents additional requirements and timing for the cohort. All diagnosis codes and pharmacy fills from claims in the 6 months prior to new SGLT-2 start were used. To reduce the high dimensions of diagnoses and pharmacy fills to an amenable size for machine learning algorithms, we grouped the diagnoses using the Hierarchical Chronic Conditions (HCCs, used by Medicare for risk adjusting total costs), and grouped the medications using AHFS Pharmacologic-Therapeutic Classifications (both dichotomous and total fills). Patients in the cohort were classified as cases (n=1237) based on having one or more diagnostic or procedural codes identifying one or more non-traumatic, lower extremity amputation following the SGLT-2 first fill date. From the universe of the eligible control cohort, we sampled 9 control patients for each case patient, matching on age, gender and race, yielding 11,133 total control patients. Bigger control size was kept to retain control heterogeneity. Table 1 shows the well-matched summary statistics of the case and control cohorts. Although the race variable in our dataset is imputed based on probability conditioned on last name and geography and is likely prone to error2, they are only used to match case and control cohorts.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Amputation (Cases)</th>
<th>No Amputation (Controls)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (mean, years)</td>
<td>58.5 (SD 10.6)</td>
<td>58.5 (SD 10.7)</td>
</tr>
<tr>
<td>Gender - N, %</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>380</td>
<td>3,420</td>
</tr>
<tr>
<td>Male</td>
<td>857</td>
<td>7,713</td>
</tr>
<tr>
<td>Race/Ethnicity - N, %</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Black</td>
<td>122</td>
<td>1,098</td>
</tr>
<tr>
<td>White</td>
<td>844</td>
<td>7,596</td>
</tr>
<tr>
<td>Hispanic</td>
<td>185</td>
<td>1,665</td>
</tr>
<tr>
<td>Unknown</td>
<td>86</td>
<td>774</td>
</tr>
</tbody>
</table>

1451
Thus, the imputation error will not disproportionately affect case/control cohort, mitigating concerns over such errors.

**Machine learning:** We experimented with the following machine learning models that allow read off of linear coefficients for easy interpretation: penalized logistic regression (i.e., ridge regression), support vector machine (SVM) with l1, and l2 regularizations respectively. Many HCC and American Hospital Formulary Service (AHFS) features are sparse. Following common machine learning practice, we retained features that are non-zero in at least 1% of the training patients, which yields 136 features. We split the combined case and control dataset into a training set, validation set, and test set by stratified splitting the samples using a 6:2:2 ratio. We trained the model using the training set and tuned the parameters using the validation set.

**Experiments and Results**

The validation set favors the l2 regularized ridge regression as the classification algorithm. **Figure 2** shows that this algorithm achieves a competitive AUC of 0.853 on the held-out test set, indicating strong predictive power of current features on lower extremity amputation risks. **Table 2** shows the top features associated with amputation risks in T2D patients taking SGLT2. The results suggest that recent skin, bone, neurological, vascular conditions, and chronic complications from diabetes are predictive of risk of amputation. Medications such as basal insulins, and platelet-aggregation inhibitors are associated with risks of amputation, possibly by identifying high risk individuals (i.e. those with more advanced diabetes and those with atherosclerotic vascular diseases). Lincomycins are used to treat resistant bacterial infections, which may identify patients with poorly healing diabetic foot ulcerations. It is also not surprising that evidence for prior amputation is a strong predictor of later amputation.

![Figure 2. ROC Curve for l2 regularized ridge regression on the held-out test set.](image)

<table>
<thead>
<tr>
<th>Feature</th>
<th>Weight</th>
<th>Feature</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chronic Ulcer of Skin, Except Pressure</td>
<td>0.231</td>
<td>Diabetes w/ Chronic Complications</td>
<td>0.074</td>
</tr>
<tr>
<td>Polyneuropathy</td>
<td>0.131</td>
<td>Vascular Disease</td>
<td>0.072</td>
</tr>
<tr>
<td>Bone/Joint/Muscle Infections/Necrosis</td>
<td>0.128</td>
<td>Lincomycins</td>
<td>0.063</td>
</tr>
<tr>
<td>Prior Lower Limb Amputation</td>
<td>0.104</td>
<td>Basal Insulins</td>
<td>0.054</td>
</tr>
<tr>
<td>Diabetic Retinopathy/Vitreous Hemorrhage</td>
<td>0.078</td>
<td>Platelet-Aggregation Inhibitors</td>
<td>0.051</td>
</tr>
</tbody>
</table>

**Discussions and Conclusions**

In this work, we demonstrated the feasibility of using machine learning models on insurance claims data to predict amputation for T2D patients starting an SGLT-2 medication. To this end, we focused on models that are easy to interpret and features broadly applicable and obtainable from an administrative database. While many known factors greatly contribute to the predictions, the inclusion and testing of other factors such as skin, bone and neurologic conditions can help improve risk prediction and be used by clinicians to make informed decisions for prescribing medications for their diabetic patients. Although the safety alert warns against prescribing of SGLT-2s for high risk patients, there may be group of high risk patients who can benefit from SGLT-2 use but show minimal increased risk of amputation. The initial prediction only used broad diagnosis and medication categories and testing of specific codes and medications could improve the predictive power and inform more targeted interventions.

**References**

Patient Preferences for Accessing, Communicating, and Sharing Health Information using Visualized Patient-Reported Outcomes

Sabrina Mangal, PhD, RN1, Leslie Park, BS1, Meghan Reading Turchioe, PhD, MPH, RN1, Lisa Grossman Liu, PhD2, Annie C. Myers, MA1, Brittany Taylor, BS, RN3, Parag Goyal, MD, MSc4, Lydia Dugdale, MD, MAR5, Ruth M. Masterson Creber, PhD, MSc, RN1

1Department of Population Health Sciences, Division of Health Informatics, Weill Cornell Medicine, New York, NY; 2Department of Biomedical Informatics, College of Physicians and Surgeons, Columbia University, New York, NY; 3Columbia University School of Nursing, New York, NY; 4Department of Medicine, Divisions of Cardiology and General Internal Medicine, Weill Cornell Medicine, New York, NY; 5Department of Internal Medicine, Columbia University, New York, NY

Introduction
Patient-reported outcomes (PROs) are a common and growing method to collect symptom information that help to inform patients’ health statuses and support their self-management. Recent research has integrated the collection and analysis of PROs to inform the development of interventions and frame patients’ clinical pictures. Moreover, with an increasing emphasis on patient engagement in care and patient-centeredness, it is also important that collected PRO information is returned to patients in an understandable format to facilitate patient-provider communication. However, there is a dearth of research as to the most effective way to deliver PRO information back to patients. To fully understand the utility and implications of returning PRO information to patients, we interviewed participants with heart failure about their experiences with a visualized symptom report and explored implications for future use.

Methods
Patients with heart failure were included in this study (diagnosis confirmed via the patient’s electronic health record) which began in February 2020. Interviews were conducted at the completion of the study where heart failure patients were periodically shown a visually enhanced symptom report webpage after each completion of symptom surveys bi-weekly for 8 weeks (Figure 1). Symptom reports were previously developed using user-centered design methods and integrate visual analogies into the information display. The reports summarized overall physical and mental health, as well as specific symptoms: depression, anxiety, pain, sleep trouble, fatigue, anger, and brain fog. After completing the study, we spoke to patients to determine the usefulness of this data, suggestions for improvement, and explored their levels of trust and preferences for sharing and receiving health information for future research studies. Three trained researchers conducted interviews with participants using a semi-structured qualitative interview guide. Interviews were audio-recorded and transcribed. Researchers used Dedoose to organize coding and content analysis of the transcripts and established inter-rater reliability across three researchers. Two researchers met weekly to compare coding progress, consulting a third researcher if consensus was needed.

Results
To date, 16 participants were interviewed, and 50 participants completed at least one symptom survey. Data collection will continue until data saturation is reached. Qualitative participants (n = 16) had an average age of 63.6 years and were 81% male, 75% White, and 19% Hispanic/Latino. A majority of the participants (94%) reported having an associate degree or higher.

<table>
<thead>
<tr>
<th>Emerging Theme</th>
<th>Exemplar Quote(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Category</td>
<td></td>
</tr>
</tbody>
</table>
Reaction to receiving data/usefulness:

<table>
<thead>
<tr>
<th>Participants stated that returning their health information and openness should be the default of clinical encounters.</th>
<th>Yeah, I think it’s a good practice. If patients don’t want to hear it, they can make a negotiation with their doctor about it. But I think it’s a good idea that you’re open, and show them, and demonstrate if there’s a change in progress that you detect. –P19</th>
</tr>
</thead>
<tbody>
<tr>
<td>The visualized symptom report served as a communication prompt between patients and clinicians.</td>
<td>That’s a basis for talking about expectations potentially, if you change medicines or change the treatment. –P20</td>
</tr>
<tr>
<td>It also gives patients a sense of their doctor’s view of how they’re doing. And sometimes, … the view[s] may not be in sync … [which] allows you to bring that up to the fore for discussion with them. –P19</td>
<td></td>
</tr>
<tr>
<td>Some participants found that having access to their own health information relieved them of the pressure of a clinical encounter.</td>
<td>[The visualization is] something tangible that I can look at and try to figure out on my own without being under pressure sitting in front of a doctor having that little amount of time. –P21</td>
</tr>
</tbody>
</table>

Reaction to content:

<table>
<thead>
<tr>
<th>Overall, participants found the symptom information to be useful for understanding their physical and mental health symptoms.</th>
<th>I still have the survey and it would be like ‘your mental health is poor’. …I hadn’t really thought about it enough until I took the survey. I was grateful they included more than just cardiac symptoms. It forced me to look at that and think maybe I should de-stress. –P25</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participants enjoyed seeing a visual analogy to indicate their progress.</td>
<td>You look at the smiley face. If it’s smiling…hopefully it’s smiling…some days when I have an off week, it lets me know how far I am off… So, it’s good feedback. It keeps me positive and lets me know what I need to work on. –P24</td>
</tr>
</tbody>
</table>

Suggestions for improvement:

| Participants wanted a mechanism for further explaining mental health symptoms based on external circumstances. | Sometimes the depression area I would look at a little more because it was also right before COVID. I think there was a lot of anxiety going on, too. –P14 |

Data sharing implications and preferences:

| Most participants would primarily use the symptom reports for themselves and would share with a clinician if needed. | If there was something significant, I would share it with my physician, but mostly it would be to inform my own choices and my own decisions. –P25 |
| Transparency was critical for patients when considering sharing their health information with different groups such as health-tech companies. | I’d have to look into it to make sure it was all randomized and private… I would be comfortable with it if I looked into it more. –P25 |

Conclusions

Overall, visualized symptom reports using visual analogies were useful to heart failure patients, which is notably different from prior research that encourages the use of line graphs. These symptom reports provided a way to prompt conversations with clinicians and help patients further understand their health status independently. Future research should consider returning health information to patients in a comprehensible format using visual aids, and should allow patients with opportunities to provide additional context to their symptom reports to communicate to clinicians.

References


Research supported by: National Institute of Nursing Research. (R00NR016275; R00NR016275-05S1; PI: Masterson Creber)
A Deep Learning Framework Using a Pre-trained BERT Model to Predict the Risk of Progression from Mild Cognitive Impairment to Alzheimer's Disease

Chengsheng Mao, PhD1, Jie Xu, PhD2, Luke Rasmussen, MS1, Jennifer Pacheco, MS1, Guoqian Jiang, MD, PhD3, Fei Wang, PhD2, Richard Isaacson, MD2, Jyotishman Pathak, PhD3, Yuan Luo, PhD1
1Northwestern University, Chicago, IL; 2Weill Cornell Medicine, New York, NY; 3Mayo Clinic, Rochester, MN

Introduction

Alzheimer’s disease (AD) is a progressive neurological disorder that begins with an early symptomatic pre-dementia stage called mild cognitive impairment (MCI). Predicting the risk of MCI-to-AD progression is important for clinical prognostication, risk stratification and early intervention. Few studies have used Electronic Health Records (EHRs) for MCI-to-AD progression prediction. Compared to the previous approaches based on biomarkers like cerebrospinal fluid (CSF) and MRI, routinely collected clinical data from EHRs will reflect real-world evidence and will be a less invasive source. In general, clinical notes in EHRs contain rich information, such as family history, laboratory values, Mini-Mental State Examination (MMSE) scores, among many other details, which may suggest disease progression. On the other hand, the Bidirectional Encoder Representations from Transformers (BERT) model has shown promising results in many NLP tasks, and has been extended to BioBERT2 and Bio+Clinical BERT3 (BC-BERT) for biomedical text and clinical narratives, respectively. However, clinical notes related to MCI and AD usually have some differences from general clinical texts in linguistic characteristics, motivating the need for specific BERT models for MCI-to-AD prediction. In this study, we developed a deep learning framework based on BERT for MCI-to-AD risk prediction using clinical notes from EHRs, and validated it on an independent dataset from a different institute. In addition, we designed a stratified batch sampler to address the class imbalance issue between case and control.

Methods

Study cohort. We identified a cohort of patients with MCI from the Northwestern Medicine Enterprise Data Warehouse (NMEDW) and Weill Cornell Medicine (WCM) using ICD-9 (331.83) and ICD-10 (G31.84) codes. Those who progressed to AD (identified by ICD-9 (331.0) and ICD-10 (G30.9)) are considered as the case group, and the control group is MCI patients who have not yet been diagnosed with AD. We identified 396 cases (average age 76.76, 58.8% females) and 3261 controls (average age 71.62, 51.5% females) from NMEDW and 435 cases and 1632 controls from WCM. Besides the above no-restrict prediction, we also considered the time windows of 6-month, 1-year, and 2-year to predict the MCI-to-AD progression in each time window, respectively. All progress notes before the first encounter when a patient was diagnosed with MCI were collected for the risk prediction model.

Framework. We first pretrained the MCI-BERT from BC-BERT on a corpus of 37K clinical progress notes with an average length of 6250 characters for MCI patients from NMEDW. We then fine-tuned MCI-BERT for the MCI-to-AD prediction task using the training patients with their notes, and evaluated it on an independent test set. In our framework depicted in Figure 1, the clinical notes of a patient were input to the pretrained MCI-BERT to generate a vector representation for the patient, then a linear layer appended with a sigmoid activation was employed to predict the probability of MCI-to-AD progression. We applied the MaxPooling strategy to generate the representation of a patient with multiple notes.

Settings. We used the default settings of BERT in pytorch transformers for MCI-BERT pretraining on MCI-related notes. For fine-tuning, the study cohort from NMEDW was randomly stratified into training and test sets by 8:2, 1/5 of the training set was set aside as a validation set used for model selection. Since the number of cases and controls was imbalanced, a small batch usually contained no case samples, making the model biased to the control class, while a large batch usually caused the out-of-memory issue for long notes. To address this issue, we designed a stratified batch sampler to ensure the same case/control ratio in all batches. In our study, the batch size was set to 4 and the max sequence length was set to 128. We used a weighted cross-entropy loss with the class weights inversely proportional to the corresponding number in a batch. In the training process, the model with the best validation performance was selected. WCM data was used to validate our framework on an independent data source.

![Figure 1 The overview of our framework](image-url)
Results and Discussion

The prediction results of our model in terms of F1 score and AUC on the test set compared to the baseline models are shown in Table 1. On NMEDW data, we observed that deep learning models (e.g., BERT-based models, LSTM and CNN) achieve higher performance than bag-of-word (BOW) model that can only achieve AUC around 0.5, indicating that only word counts in the notes cannot provide any information for MCI-to-AD prediction. Our framework (MCI-BERT) is pretrained on MCI-related notes, it could be more effective for MCI-related tasks such as MCI-to-AD prediction. MCI-BERT pretrained on NMEDW (i.e., MCI-BERT(NM)) cannot perform as well as MCI-BERT(WCM) on WCM dataset, which is due to the different note language styles between the two institutes, suggesting the importance and difficulty of developing a cross-institute pretrained language model. The significant performance difference between NMEDW and WCM by the same model also indicates the discrepancy between the two datasets. In the previous reports, Young et al.4 and Davatzikos et al.5 achieved an AUC of 0.795 and 0.734, respectively, using biomarkers in their risk prediction models. In comparison to these studies, besides clinical notes are more routinely available from large samples of patients than previous biomarkers, our performances also validate the effectiveness of using clinical notes for MCI-to-AD prediction. Figure 2 shows the attention of each word that MCI-BERT relies on to make the prediction, where MCI-BERT pays more attention to the term “memory loss” than others, which is reasonable as serious memory loss is a symptom of AD.

![Figure 2](image)

**Figure 2** Attention visualization of MCI-BERT

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Model</th>
<th>No-restrict</th>
<th>6-month</th>
<th>1-year</th>
<th>2-year</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F1</td>
<td>AUC</td>
<td>F1</td>
<td>AUC</td>
<td>F1</td>
</tr>
<tr>
<td>NMEDW</td>
<td>MCI-BERT(NM)</td>
<td>0.3753</td>
<td>0.8247</td>
<td>0.3578</td>
<td>0.7395</td>
</tr>
<tr>
<td></td>
<td>BERT-base1</td>
<td>0.2669</td>
<td>0.7223</td>
<td>0.3117</td>
<td>0.7033</td>
</tr>
<tr>
<td></td>
<td>BioBERT2</td>
<td>0.3899</td>
<td>0.8039</td>
<td>0.3179</td>
<td>0.7251</td>
</tr>
<tr>
<td></td>
<td>BC-BERT3</td>
<td>0.3163</td>
<td>0.7109</td>
<td>0.352</td>
<td>0.7233</td>
</tr>
<tr>
<td></td>
<td>BiLSTM</td>
<td>0.3258</td>
<td>0.7819</td>
<td>0.2885</td>
<td>0.6069</td>
</tr>
<tr>
<td></td>
<td>BiLSTMatt</td>
<td>0.3459</td>
<td>0.8021</td>
<td>0.2727</td>
<td>0.621</td>
</tr>
<tr>
<td></td>
<td>CNN</td>
<td>0.3605</td>
<td>0.7969</td>
<td>0.2857</td>
<td>0.6592</td>
</tr>
<tr>
<td></td>
<td>BOW+LR</td>
<td>0.1805</td>
<td>0.5004</td>
<td>0.1698</td>
<td>0.4765</td>
</tr>
<tr>
<td>WCM</td>
<td>MCI-BERT(WCM)</td>
<td>0.5225</td>
<td>0.7896</td>
<td>0.1714</td>
<td>0.6246</td>
</tr>
<tr>
<td></td>
<td>MCI-BERT(NM)</td>
<td>0.4661</td>
<td>0.7598</td>
<td>0.2000</td>
<td>0.6062</td>
</tr>
<tr>
<td></td>
<td>BERT-base1</td>
<td>0.4400</td>
<td>0.7348</td>
<td>0.1395</td>
<td>0.5932</td>
</tr>
<tr>
<td></td>
<td>BioBERT</td>
<td>0.4729</td>
<td>0.7592</td>
<td>0.1848</td>
<td>0.5575</td>
</tr>
<tr>
<td></td>
<td>BC-BERT</td>
<td>0.4674</td>
<td>0.7730</td>
<td>0.1333</td>
<td>0.5613</td>
</tr>
</tbody>
</table>

Conclusion

Clinical notes contain rich information that may suggest disease progression. However, it is challenging to extract predictive information from unstructured notes. The deep learning framework using BERT models in this study may provide a solution for clinical note analysis for MCI-to-AD prediction. Our study also demonstrates the discrepancy of clinical note styles between different institutes.

References

Integrating Evaluations of Predictive Algorithm-Driven Interventions into Clinical Workflows with the Dynamic Discontinuity Deployment Design

Ben J. Marafino, BS¹, Alejandro Schuler, PhD¹,², Vincent X. Liu, MD, MS³, Art B. Owen, PhD¹, Gabriel J. Escobar, MD³, Mike Baiocchi, PhD¹

¹Stanford University, Stanford, CA, USA; ²unlearn.ai, San Francisco, CA, USA; ³Division of Research, Kaiser Permanente, Oakland, CA, USA

Introduction

Electronic health record (EHR) systems now increasingly incorporate predictive algorithms and display their output at the point of care as a form of clinical decision support. This output often takes on the form of an outcome risk score and in some cases, can be used to initiate an intervention or care pathway with the goal of improving a certain clinical outcome. For example, if a patient’s predicted risk of 30-day readmission exceeds a certain threshold value at hospital discharge, then they are automatically enrolled in a care coordination pathway designed to reduce their readmission risk—a process which occurs entirely within the EHR. However, in the absence of some type of formal evaluation, it is unclear whether deployments of these predictive algorithm-driven interventions (or “prediction-action dyads”) actually improve clinical outcomes.¹

Randomized clinical trials (RCTs) represent an ideal choice with which to evaluate the impacts of these deployments on outcomes. However, such RCTs are uncommon, in large part due to the burdens that conducting these trials may impose on existing clinical workflows. In addition, randomization schemes must be carefully designed to avoid unintentional bias, which can be infeasible—for example, if randomizing at the unit of the care team rather than the individual patient encounter.² Moreover, in the context of institutional oversight, many deployments of these algorithms are classed as quality improvement projects, which often precludes the use of randomization. As a result, stakeholders often must resort to uncontrolled before-after studies, which rarely yield robust evidence regarding the impacts of these interventions.

In recognizing the disadvantages of RCTs in this setting, we propose an alternative in the form of a new type of cluster-randomized trial design, which we refer to as the dynamic discontinuity deployment design, or the D3 design. Using the regression discontinuity design as a building block, the D3 design dynamically varies, across clusters, the risk score threshold used to assign a predictive algorithm-driven intervention to characterize its impact in a deployment. The D3 design more closely emulates real-world practice conditions, alleviating many of the training and workflow burdens imposed by RCTs, while also providing valuable direct evidence regarding treatment effect heterogeneity.

Methods

As proposed, the D3 design assumes the existence of a predictive model, and an intervention associated with it, wherein the intervention is assigned based on a patient’s estimated risk score relative to a predetermined score cutoff. Furthermore, the D3 design assumes I subjects, J clusters, K score cutoffs, and L time points. The J clusters may correspond to regions, hospitals, clinical units (e.g., a floor or a ward), or any type of organizational unit. The K score cutoffs, \{c₁, c₂, ..., cₖ\}, are chosen in advance and are assumed to lie on [0, 1]. The central idea of the D3 design is that at each time point 1, 2, ..., L, each cluster uses a different value of the cutoff to assign the intervention. For example, for the month of February, a hospital uses the risk score cutoff of c₃ = 25% to assign the intervention, whereas in March it uses the cutoff c₇ = 30%, and so on. This process is summarized in the left panel of Figure 1.

Thus, as time progresses, each cluster cycles between all risk score cutoffs, if L = K—or necessarily some subset of them if L < K. At the conclusion of the experiment, the data across all clusters for each of the K cutoffs is pooled and a regression discontinuity (RD) estimator is applied to estimate the local average treatment effect (LATE) \( \hat{\tau}_k \) at each cutoff k. Notably, to reduce bias, the RD estimators implicitly take advantage of the balancing property of the risk score under its interpretation as a prognostic score.³ Using these LATE estimates \( \hat{\tau}_1, ..., \hat{\tau}_K \), one can estimate and infer a pooled global treatment effect, which resembles the average treatment effect from a RCT, as well as gain insight into treatment effect heterogeneity implied by the \( \hat{\tau}_k \), as shown in the right panel of Figure 1.
To characterize the statistical properties of the D3 design, we undertake a series of simulation experiments. Our experimental approach primarily assesses the bias and consistency properties of the global and cutoff-specific (LATE) estimators, and also compares the LATE estimators to the analogous post-hoc analysis of a RCT with $K$ interaction terms on the risk scale. The heterogeneous treatment effects (HTE) are parameterized by a scaled beta density, and we assume varying extents of intra-cluster correlation (ICC).

**Results**

For both the global treatment effect, as well as the LATEs, the D3 estimator appears to be unbiased and $\sqrt{n}$-consistent, with somewhat slower convergence in the tails of the risk score distribution. These results appeared robust to various choices of the simulation parameters, including the choice of the HTE parameterization as well as when the ICC was varied. Moreover, compared to the post-hoc analysis of the analogous RCT using risk $\times$ treatment interaction terms, the D3 estimates of the LATEs were 10% to 30% more efficient than the interaction estimates (i.e., $\hat{V}_{\text{D3},k}/\hat{V}_{\text{RCT},k} \approx 0.7$ to 0.9, where $\hat{V}_{\cdot,k}$ represents the variance of either method at or around the cutoff $k$.)

**Discussion**

As verified in simulation experiments, the D3 design appears to have desirable statistical properties. In particular, when compared to the analogous randomized controlled trial, the D3 design incurred a lower cost in terms of variance in its HTE estimates when comparing the cutoff-wise treatment effects to the corresponding interaction estimates. Moreover, the D3 design has another advantage over a RCT in that it may reduce the “friction” that evaluation processes often introduce into clinical workflows. Altogether, the D3 design may be a feasible choice for organizations and health systems considering methodologies to evaluate deployments of predictive algorithm-driven interventions. Further work remains, including fully characterizing the qualitative and process aspects of the D3 design, determining which cutoff schedules may be optimal, and whether the design could be made more efficient, either via borrowing data from other cutoffs or by introducing adaptivity.

**References**

Managing Health Care Knowledge in Predictive Modeling, Clinical Decision Support and Metric Reporting – eXecutable Library Architecture (XLA)

Susana B. Martins, MD MSc, Cora L. Bernard, PhD MS, Suzanne Tamang, Jodie A. Trafton, Ph.D., M.S.
VA Program Evaluation and Resource Center, VA Office of Mental Health and Suicide Prevention (OMHSP), Menlo Park, CA; Department of Biomedical Data Science, Stanford University, Stanford, CA, USA; Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford, CA, USA.

Introduction
Face-validity, consistency and accuracy of clinical concepts utilized in Clinical Decision Support Systems (CDSS) are paramount for effective use and acceptance. While standard vocabularies have been developed to support knowledge translation, in reality, capture of health care information is locally nuanced and shaped by, for example, Electronic Health Record (EHR) design features such as clinical prompts and documentation tools (e.g., pick lists, note templates, etc.), clinical workflow, and local culture and practice in coding. For use in CDSS, data definitions for clinical concepts must be maintained in alignment not only with an evolving medical knowledge base, but also with local practice patterns and EHR systems. For each clinical concept within each unique EHR implementation, this requires understanding both what to look for and where to find it in downstream databases. Accurate identification of clinical concepts requires not only specification of value sets of clinical vocabulary, but also of the data locations and logic for combining data elements, information that is rarely captured in clinical data libraries. This information is typically captured only in code, limiting transparency and reusability of validated data definitions.

Here, we provide a solution for integration of clinical concept libraries directly within deployed data systems. The eXecutable Library Architecture (XLA) provides a data-driven approach to integrate value set and clinical concept metadata along with best practices in data extraction, enabling maintenance of a library of validated clinical concepts that can be flexibly applied and directly executed to produce datasets required for predictive modeling, CDSS functions or quality and performance measurement, at national, facility or clinician level.

Methods
We created a standard querying language (SQL)-based architecture that seamlessly integrates metadata with code, eliminating the implementation gap between defining and computing complex clinical concepts. The architecture contains two schemas: a portable and Protected Health Information (PHI)-naïve Library and a data-driven Dataset.

The Library has a modular design for transparency and efficiency. The value set functions as the building block and can be grouped into supersets and logically attached (e.g., AND, OR, NOT) or temporarily related (e.g., WITHIN 8 Days) to other sets to create arbitrarily complex concepts. Implementation instructions at each level of aggregation are encoded with exportable and visualizable metadata (Figure 1). The Library definition of a concept is the sole source of instructions for its computation. The system allows end-users to enter requests for concepts and associated time frames of interest (e.g., 1 year prior to selection date) and the architecture automatically generates the dataset. Specifically, the Dataset schema will populate and execute dynamic SQL code to extract the necessary value sets over the desired time frames from the appropriate data sources, seamlessly combining multiple sources such as VistA and Cerner-Millennium sourced tables while tracking the originating context, such as outpatient or inpatient diagnoses. Following value set extraction, the Dataset construction code rolls up the raw data into the requested concepts according to the Library instructions, exposing its computations at each step. Code and data output are exposed to support validation. Version control is embedded in the architecture.

Results
For an initial proof of concept, we recreated the dataset underpinning the VA Stratification Tool for Opioid Risk Mitigation (STORM1), a predictive model that is used to update patient risk estimates nightly for CDS. STORM value sets include VA Office of Mental Health and Suicide Prevention national operational definitions: 360,483,480 variables for 6,675,620 patients were calculated faster (< 1 hour) and with an accuracy of 99.95% when compared to
current standard computational practice. The domains included CPT (1), ICD10 (48), Drug (6), outpatient visit types (2), and inpatient specialty (2). Metadata is seamlessly displayed in a Report (Figure 1).

**Figure 1.** XLA generates a Report on dataset metadata automatically based on data-driven computation.

**Discussion**

EHRs are a living system whose outputs are unique to their design and environment. There is a tradeoff between systems that standardize clinical data and those that support the best institutional-specific practices in extracting clinical concepts from biased data, which change rapidly and locally. To ensure accuracy, XLA pairs customizable knowledge management tools to define variables with the generation of SQL code that searches the VA’s data warehouse to retrieve relevant items that belong to each variable’s value set. Other knowledge engineering frameworks relevant to the deployment of predictive analytics in healthcare that build on open standards include SOLOR and OMOP Common Data Model. However, they focus on the knowledge engineering aspect and do not easily facilitate the execution of CDSS. Although they can be customized to exploit an institution’s data assets and information systems, these tools are designed to support generalizable frameworks for data sharing and interoperability across disparate health systems. XLA is transparent and intuitive to use for SMEs. In addition, it generates and executes code specific to an institutional data warehouse for the purpose of real-world risk prediction as well as CDSS functions.

**XLA Advantages:**

- Provides standardization of clinical concept definitions & methods for computing from source databases.
- Enables reuse of clinical concepts across analytic endeavors: CDSS, predictive modeling, metrics, etc.
- Organizes and documents data definitions in a clear format for knowledge management.
- Transparently computes clinical concepts, making them usable to a wide variety of stakeholders.

Future work will focus on optimization for increased computational efficiency, strengthening the validation and embedded version control features, and expanding the metadata report.

**Conclusion**

The XLA architecture successfully computed a clinically relevant dataset with a high level of accuracy and performance. Integrated metadata for this dataset were displayed in a report. We are currently implementing this architecture to support predictive modeling, CDSS and metric reporting.

**References**


Acknowledgements: The authors wish to thank Shalini Gupta, Dan Hardan, Amy Robinson, Rebecca Stephens, and Elizabeth Oliva for their contributions to the XLA architecture.
Longitudinal Performance of a Machine Learning NICU Sepsis Predictor

Aaron J. Masino, PhD\(^1\)\(^2\), Mary Catherine Harris, MD\(^1\)\(^2\), Lakshmi Srinivasan, MD\(^1\)\(^2\), Robert W. Grundmeier, MD\(^1\)\(^2\)

\(^1\)Children’s Hospital of Philadelphia, Philadelphia, PA; \(^2\)University of Pennsylvania Perelman School of Medicine, Philadelphia, PA

Introduction

Despite the importance of early treatment, timely recognition of sepsis remains difficult among infants hospitalized in neonatal intensive care units (NICUs) due to heterogeneous and often indolent presentation. A number of predictive models have been proposed to automate early sepsis recognition in both adult and pediatric settings using a variety of methods including machine learning\(^1\). Nearly all of these studies present model performance results for a single point in time, typically the time of blood culture order. However, an operational sepsis early warning system (EWS) will necessarily make multiple predictions over time during which variation in data availability and clinical status will affect model performance. For example, the median length of stay of infants hospitalized in the Children’s Hospital of Philadelphia (CHOP) NICU is 12 days (IQR [5, 34]). Here, we examine the longitudinal performance of a machine learning sepsis prediction model over a 24-hour period for a NICU setting.

Methods

This study was approved by the CHOP IRB. We conducted a retrospective analysis of infants hospitalized in the CHOP NICU between January 2013 and September 2019. We obtained electronic health record (EHR) data for all infants with at least one sepsis evaluation. The 922 infants in our dataset had 847 positive sepsis evaluations. We obtained case data from the 48-hour period ending at the time of clinical sepsis evaluation, defined as the time of blood-draw for culture or initiation of antibiotic treatment, whichever was earlier, for evaluations resulting in positive cultures for bacteremia or negative cultures with ≥120 hours of antibiotic treatment (commonly described as “clinical sepsis”). We obtained 1,412 control samples by randomly selecting 48-hour observation periods from all individuals in our NICU sepsis registry during periods for which there was no evidence of sepsis for ≥10 days before or after the observation period. For model development, we used 35 features including vital signs, demographics, lab measurements, chronic conditions, and clinical observations. We also used threshold and vital sign difference features, where we computed the difference between the current value and the mean over the previous 24 hours. We binned the data into 48 one-hour windows such that the final bin, denoted bin 1, ends at the time of sepsis evaluation for cases with data bin assignment based on its EHR timestamp. Due to high missingness rates, lab values were treated as indicator variables (i.e., the lab feature was set to 1 if the lab result was available in the bin hour, 0 otherwise). All other features were carried forward in time if there was an observation earlier in 48-hour time period to address missing values. Otherwise, we used mean imputation. We used the first 24 hours, bins [48, 25], of the 48-hour period to populate the vital sign difference features and carry forward values into the final 24 hours, bins [24, 1]. We trained an XGBoost\(^2\) classification model to predict sepsis probability using a leave-one-out cross validation (LOO-CV) approach using only the data available in bin 1 (the hour ending with sepsis evaluation). For each LOO-CV iteration, the model was used to predict sepsis for each hour in the last 24 hours of the held-out episode, bins [24, 1]. We selected the decision threshold to achieve 80% sensitivity in the last hour, the end of which corresponds to the sepsis evaluation time for cases. We assessed model performance by calculating sensitivity, positive predictive value (PPV), and area under the curve (AUC) of the receiver operating characteristic over all predications in each of the last 24 hours.

Results

We found that model performance was nearly constant over the first 20 hours, followed by rapid improvement in all performance metrics during the final 4 hours (left, Figure 1). We also found that the mean probability of sepsis for cases was more than double that of controls at each hour (right, Figure 1). Furthermore, while the mean probability remained constant over the entire 24-hour period for the controls, there is an apparent linear increase over the first 20 hours followed by a more rapid, non-linear increase in the final 4 hours for sepsis cases. To gain further insight into potential influences on model performance, we examined the update frequency of each of the model input features (Figure 2). Nearly half of the input features, including all lab values, were updated less than once per day, implying most information available to the model does not change during the 24-hour period. Kernel density estimates (not shown), suggest however, that in the case samples there is a bias toward lab value observation in the 4 hours prior to sepsis evaluation, whereas the distribution of hour of observation is approximately uniform for controls.
Figure 1: (Left) Model performance metrics at each hour. (Right) The mean probability of sepsis at each hour across all cases (blue, squares) and controls (red, circles). The shaded regions indicate standard deviation.

Figure 2: Mean updates per hour for longitudinal varying input features (excludes chronic conditions which we considered to be a baseline input for the entire period and derived vital sign difference variables). Error bars indicate standard deviation. The dashed line corresponds to an update frequency of once per day. MAP: mean arterial pressure. CVL: central venous line. UAC: umbilical arterial catheter. ITRATIO: neutrophil immaturity ratio.

Discussion

The consistent model performance observed over time suggests the model is stable and may be useful in screening for sepsis in a NICU setting. However, we note that at about 80% sensitivity, the model sustained about 60% PPV which may lead to an unacceptable number of sepsis evaluations per infant depending on resource availability. The difference in mean probability of sepsis between cases and controls over the entire evaluation period suggests the difference in baseline sepsis risk is discernable at least 24-hours prior to clinical recognition. The observation that the mean probability of sepsis for controls is constant over the entire period is suggestive that a combination of relatively static features (i.e., gestational age, comorbidities, presence of central lines) establishes the baseline risk. Noting that we used a carry forward approach for missing values, most episodes have a complete set of recent observations during the final hours of the window. Thus, if the dynamic variables establish baseline risk, we would expect a change in sepsis probability for controls toward the end of the 24-hour period. The increase in mean sepsis probability for cases and improved model performance at the end of the 24-hour period indicate some change in input features that improves classification. It is likely that evolving sepsis produces changes in many, or all, of the variables that improve classification only a few hours prior to clinician recognition. However, we noted there was a bias toward updated lab observations in the final 4 hours for cases. Since we treated labs as indicator variables (i.e., the model does not use the actual lab result only the fact that the lab was observed), it may be that the observation of a lab value is a proxy for clinician suspicion of deterioration. We are continuing to investigate these questions through our ongoing research.

References

“There for You Every Day”:
Pilot Study of an Automated Conversational Health Coach

Elliot G Mitchell, MA, MPhil, Lena Mamykina, PhD

1Columbia University, Department of Biomedical Informatics, New York, NY

Introduction

In-person self-management education and coaching are among the most successful interventions for managing chronic conditions. In particular, health and lifestyle coaching aims to help individuals meet personal health goals through interaction with a health coach, with an emphasis on developing essential skills and knowledge, cultivating motivation, and personalizing support for each individual. Unfortunately, many individuals, especially those from disadvantaged communities, face barriers to accessing coaching support, and there are not enough coaching practitioners to reach the growing population living with chronic diseases. Because coaching takes the form of a conversation between a coach and a client, conversational agents are well-positioned to deliver coaching interventions, and can reach broader and more diverse populations than their in-person counterparts. However, key questions remain in designing effective conversational agents for health coaching.

A common approach to designing conversational agents in health is structuring dialogs to follow clinical protocols for specific coaching strategies, like Cognitive Behavioral Therapy. In this approach, designers create a scripted dialog structure that anticipates the possible interactions between an individual and the agent and specifies appropriate responses. While less flexible than dynamic conversational agents, which are based on dialog models trained from a large corpus, fully-scripted agents have been successful in domains where corpora are not available or feasible to create, and also offer designers more precise control over how the agent will respond, which is important when delivering health-related information. Despite their potential to serve the role of a health coach, conversational agents have not been widely studied in their ability to support chronic disease self-management.

Methods

We designed t2.coach through an iterative, user-centered design process. Following the dominant approach for designing conversational agents in health, t2.coach was implemented as a scripted chatbot following an established protocol for goal-setting, Brief Action Planning (BAP), which lays out a series of steps and scripted prompts for practitioners to facilitate goal setting conversations. Inspired by BAP, interactions with t2.coach start by guiding the user through a series of prompts to choose a health goal related to nutrition or physical activity and plan specific actions to achieve that goal. Every day, t2.coach initiates a check-in conversation to inquire whether the user met their goal, and if they didn’t, what barriers they encountered and how they might overcome them; users can choose a new plan each day. See Figure 1 for an example exchange.

t2.coach was implemented as a text message (sms) based chatbot with a botkit backend and twilio integration for sms messaging. Similar to Woebot and other health chatbots, users could reply with a mix of multi-response, yes/no, and free-text prompts, though the free text responses did not impact the overall flow of the conversation.

We sought to examine the feasibility of automated conversational approaches like t2.coach to serve as a health coach and support individuals with type 2 diabetes (T2D). In particular, we assessed the extent to which individuals engaged with t2.coach, and captured qualitative perceptions from individuals after using t2.coach for 2-3 weeks.
To examine these questions, we recruited individuals with T2D from community health centers in the New York City metro area as well as from Columbia University’s online RecruitMe portal to participate in a series of focus groups and pilot studies. In the first phase, participants messaged with an early “wizard-of-oz” (woz) prototype, where messages were sent by a member of the research team. We updated the t2.coach script based on feedback from the woz pilot, and in the second phase, participants interacted with a functioning prototype of t2.coach. Both pilot studies ran for 2-3 weeks, followed by semi-structured interviews with participants. We calculated descriptive usage statistics of interactions from both pilots. Interview transcripts were analyzed with inductive thematic analysis.

Results

20 individuals with T2D participated in the pilot. Of those, 13 were enrolled in the woz pilot, while 7 were enrolled in the pilot with the functioning chatbot.

Iterative Design. There were several findings from the woz pilot that led to updates in the t2.coach script. For example, we found that the full BAP script, which was designed for an in-person session, was sometimes prohibitively long for text messaging, and shortened it to make it more concise. We also updated and expanded the available content – goals and action plans to choose from based on focus group feedback.

Engagement. Participants responded to daily messages 50% of the time with the woz prototype, and 70% of the time with the fully implemented t2.coach chatbot. This difference was statistically significant with an unpaired t-test ($t < 0.05$), and suggests that the modifications made through the iterative design process improved the user experience for participants. When users did respond, users took an average of 16.9 minutes to reply suggesting a substantial delay compared to synchronous human-human conversations.

Qualitative Perceptions. Thematic analysis across both pilot studies identified several main themes, summarized here:

- **Working with a coach.** Overall, participants described experiences as consistent with working with a health coach. Participants used words like “coach” or “teacher” to describe the system. Through setting and following goals, participants described increased motivation, and a sense of accountability to achieve their chosen goals. Participants also commented that their exchanges with the coach felt like a conversation, and that they felt they were building a relationship with t2.coach through their messaging.

- **Annoying, but helpful.** Some participants described the frequency of messages as “annoying,” especially when messages would arrive at inopportune times. This was particularly the case for participants who used text messaging infrequently prior to the study. At the same time, participants often found t2.coach’s persistence to be helpful, as it provided necessary encouragement for following through on behavioral intentions.

- **There for you every day.** Many participants had prior experience with a health coach, like a dietitian, who they saw every few months, and whose recommendations often took the form of a meal plan. Participants contrasted t2.coach with these prior experiences, and appreciated the flexibility and autonomy of setting goals with t2.coach, as opposed to the prescriptive and rigid nature of meal plans. Furthermore, the daily messaging from t2.coach helped to keep up motivation, and the ability to adjust their daily plans to meet goals helped to avoid monotony.

Discussion and Conclusion

We presented the iterative design and a multi-stage pilot trial of t2.coach, a conversational agent health coach designed to support individuals with T2D. We refined t2.coach with a wizard-of-oz trial, followed by a pilot deployment of the functional chatbot. Qualitative results suggested that a scripted, rule-based chatbot can lead to engagement and cultivate a coach-like relationship in the context of self-management. Future work should explore the potential of more advanced computational approaches to facilitate improved automated coaching conversations.

References

Pragmatic Patient Reported Outcomes via EHR Patient Portal vs. Telephone: A Pilot Randomized Controlled Trial among Adults with Epilepsy and Anxiety or Depression Symptoms

Heidi M. Munger Clary, MD, MPH, Beverly M. Snively, PhD, Umit Topaloglu, PhD, FAMIA, Pamela Duncan, PhD, James Kimball, MD, Halley Alexander, MD, Gretchen A. Brenes, PhD

1Wake Forest School of Medicine, Winston-Salem, North Carolina

Abstract

This pragmatic pilot randomized trial was designed to assess feasibility of patient reported outcome collection at 6-months via the electronic health record (EHR) patient portal versus telephone among adults with epilepsy and anxiety or depression symptoms. While retention was lower in the EHR arm than telephone (67% vs. 100%), EHR collection required significantly less staff time. A hybrid of these two approaches may be optimal in future studies.

Introduction

To advance pragmatic trials and other research in routine care settings, study recruitment at the point of care using integrated EHR functionality may help identify and recruit more generalizable study samples. In addition, EHR portals may help facilitate rigorous outcome collection with low participant burden. Studying EHR-based clinical research methods among people with comorbid anxiety or depression and epilepsy may have high impact1, notwithstanding these comorbidities are understudied. Despite patient interest in research participation, even a single in-person research visit can be a significant deterrent to treatment study enrollment2-3. Thus, the objective was to assess feasibility of 6-month patient reported outcome collection and examine process measures in a randomized study of EHR patient portal versus telephone interview outcome methods among adults with epilepsy and anxiety or depression symptoms.

Methods

Recruitment began at a tertiary epilepsy clinic visit, with initial EHR-embedded screening consent and automated EHR eligibility assessment (Figure 1). Quality measure-satisfying anxiety and depression screening was conducted at clinic visits using EHR-based Generalized Anxiety Disorder-7 (GAD-7) and Neurological Disorders Depression Inventory-Epilepsy (NDDI-E) questionnaires. Individuals with borderline or high scores (GAD-7 ≥8 or NDDI-E ≥14) immediately received a brief research screening consent within the EHR, and automated rules in the Epic® EHR were utilized to assess other eligibility criteria (age≥18, negative suicidality screen on the NDDI-E passive suicidality question). A silent Best Practice Advisory (BPA) message to the study team inbasket pool was activated for patients who were potentially eligible and interested based on the screening consent. The study team then manually assessed the EHR for the final eligibility criterion (diagnosis of epilepsy). Fully eligible and interested individuals then completed full consent, enrollment, and randomization via telephone.

Figure 1. Recruitment process and role of the EHR

Participants (N=30) were randomized 1:1 to patient portal versus telephone outcome assessment (stratified by patient portal account status at enrollment: activated vs. not), and outcomes were collected at 3- and 6-months (patient-reported outcomes: anxiety, depression, quality of life along with process outcomes). The primary outcome was 6-month retention in the patient portal arm, defined as outcomes returned via EHR portal by one week after five
electronic portal reminders (sent every 2-3 days if no outcome returned). Process outcomes included research staff time, number of reminders required, and timing of outcome collection relative to target date for outcome assessment. Staff time included time to send portal message reminders manually, conduct telephone reminder calls, and collect and enter outcome data in the study database. This initial report focuses on the retention and process outcomes.

**Results**

The study sample had a mean age of 42.5 years and was composed of mostly non-seizure free individuals (67%) with focal epilepsy (80%). Among the 15 individuals randomized to patient portal outcome assessment, 10 (67%, 95% CI 41.7-84.8%) completed 6-month outcomes by this modality, compared with 100% (95% CI 79.6%-100%) of those randomized to the telephone interview (p=0.042). Following ascertainment of the primary retention measure, the five EHR arm participants who did not complete the outcome measures were contacted by telephone, and four completed the measures via telephone interview, while the fifth eventually returned outcome measure responses via the patient portal. Patient-portal based outcome assessment required significantly less research staff time, especially when outcome collection was successful using the EHR modality (Table 1).

**Table 1.** Process measures overall and by randomized modality*

<table>
<thead>
<tr>
<th></th>
<th>Overall (N=30)</th>
<th>EHR (N=15)</th>
<th>Phone (N=15)</th>
<th>P value**</th>
</tr>
</thead>
<tbody>
<tr>
<td>Staff time for outcome collection (minutes)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3-month</td>
<td>14.4 ± 7.3</td>
<td>11.0 ± 7.1</td>
<td>17.9 ± 5.8</td>
<td>0.0039</td>
</tr>
<tr>
<td>6-month</td>
<td>15.6 ± 9.8</td>
<td>13.5 ± 12.5</td>
<td>17.7 ± 5.7</td>
<td>0.085</td>
</tr>
<tr>
<td>Staff time for outcome collection &amp; data entry</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6-month</td>
<td>25.2 ± 14.0</td>
<td>21.4 ± 13.0</td>
<td>29.0 ± 14.3</td>
<td>0.044</td>
</tr>
<tr>
<td>Number of reminders</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3-month</td>
<td>2.8 ± 2.3</td>
<td>3.5 ± 2.5</td>
<td>2.1 ± 1.8</td>
<td>0.094</td>
</tr>
<tr>
<td>6-month</td>
<td>3.5 ± 3.3</td>
<td>4.9 ± 4.0</td>
<td>2.0 ± 1.4</td>
<td>0.068</td>
</tr>
<tr>
<td>Observations relative to due date (days)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6-month</td>
<td>3.3 ± 13.7</td>
<td>8.3 ± 17.3</td>
<td>-1.6 ± 6.2</td>
<td>0.39</td>
</tr>
</tbody>
</table>

Resources for ‘randomized modality only’ subset (those who returned outcomes via randomized modality)

|                                |                |            |              |           |
| Staff time for outcome collection (minutes) |                |            |              |           |
| 6-month (N=25, 10, 15)            | 13.0 ± 7.7     | 5.9 ± 3.6  | 17.7 ± 5.7   | <0.001    |
| Staff time for outcome collection & data entry |                |            |              |           |
| 6-month (N=25, 10, 15)            | 22.8 ± 13.6    | 13.6 ± 3.7 | 29.0 ± 14.3  | <0.0001   |
| Number of reminders              |                |            |              |           |
| 6-month (N=25, 10, 15)            | 2.3 ± 2.0      | 2.7 ± 2.7  | 2.0 ± 1.4    | 0.79      |
| Observations relative to due date (days) |                |            |              |           |
| 6-month (N=25, 10, 15)            | -1.7 ± 8.1     | -1.9 ± 10.7| -1.6 ± 6.2   | 0.34      |

*Count (column %), mean ± SD; data are complete unless otherwise noted. **P values are for comparison of EHR and telephone groups, based on Fisher’s exact and Wilcoxon rank sum tests.

**Discussion**

In this randomized pilot study, EHR portal outcome assessment had lower retention than telephone interview, but it saved significant research staff time. Patient reported measures were ultimately collected from all participants using a hybrid approach following primary outcome assessment. Future pragmatic studies may benefit from a hybrid of these approaches, should examine the relation between factors such as baseline EHR portal activity with completed EHR outcomes, and consider comparison with in-person methods. Future directions could include using automated methods to further reduce staff time for reminders. The small sample size of this study may limit its generalizability.

**References**

Predicting Level of Care for Emergency Hospital Admissions to Optimize Triage

Minh Nguyen1 M.A, Nicolai P. Ostberg2 M.S, Conor K. Corbin1 B.S, Tiffany Eulalio1 B.S, Gautam Machiraju1 B.A, Ben J. Marafino1 B.S, Michael Baiocchi3 Ph.D, Christian Rose4 M.D, Jonathan H. Chen5 M.D, Ph.D

1. Department of Biomedical Data Science, Stanford University School of Medicine, Stanford, CA 2. New York University Grossman School of Medicine, New York, NY 3. Department of Epidemiology and Population Health, Stanford University School of Medicine, Stanford, CA 4. Department of Emergency Medicine, Stanford University School of Medicine, Stanford, CA 5. Department of Hospital Medicine, Center for Biomedical Informatics Research, Stanford University School of Medicine, Stanford, CA

Introduction:
Due to the dynamic and high acuity nature of patients in the Emergency Department (ED), determining an accurate level-of-care assignment for admission is difficult. Delays in ICU admission due to under-triaging contribute significantly to mortality and increases nursing care demand [1]. Conversely, over-triaging patients may misallocate scarce ICU resources [2], which can be particularly poignant during times of high critical care stress, like during the COVID-19 pandemic.

Evidence suggests that physicians frequently make level-of-care assignments based on qualitative clinical judgment, as opposed to written guidelines [3]. As a result, these decisions may be prone to unintentional biases and other errors in judgment. Therefore, triaging patient admission assignments from the ED represents an opportunity for computer-aided clinical decision support to manage difficult triage decisions that otherwise place undue pressure on decision-makers, with potentially dire consequences for both under- or over-triaging.

In this study, we sought to develop a model to predict the level of care needed (ICU vs. non-ICU) following inpatient admission for ED patients using EHR data. Specifically, we aim to predict a patient's highest level of care within 24 hours and level of care at the 24th hours following inpatient admission, using only data available prior to the time when the admission order was written.

Methods:
We used de-identified electronic health record data for patient encounters at a tertiary academic hospital (Stanford hospital and Stanford Valley Care community hospital) and Level I trauma center between 2015 and 2019. All adult patients 18 years or older admitted to the hospital as inpatients from the ED were included. Patients who were not full code status were excluded. The primary outcome was the patient’s highest level of care within 24 hours following their initial inpatient admission. The secondary outcome was a patient’s level of care at the 24th hour since inpatient admission.

To account for feature drift, the data were split by time, with the training dataset including all patient admissions from 2015 to 2017, whereas the validation dataset included all 2018 admissions, and the test dataset included all 2019 admissions. Prediction models were trained under two regimes, corresponding to different feature sets. In Regime I, a model was constructed that used all relevant structured electronic health record data, including demographics, Emergency Severity Index (ESI), vital signs, lab results, historical diagnosis codes, and imaging, procedure, and medication orders (99,667 features). This model served as an upper bound on model performance, assuming that most structured data can be easily queried to feed the model pipeline when deployed. In Regime II, a simpler model was trained using only patient demographics and vital signs (97 features) from the current admission.

We applied four machine learning algorithms: elastic net regularized logistic regression, random forest, gradient-boosted trees, and feed-forward neural networks. Additionally, because the ESI is the triage score assigned when a patient is first assessed by triage nurses to guide treatment priority in the ED, we included a univariate logistic regression with ESI as the only predictor as a benchmark method. Predictive performance was quantified using areas under the receiver operator characteristic curve (AUROC) as well as under the precision-recall curve (AUPRC), sensitivity, specificity, positive and negative predictive values. Feature ablation studies were performed to determine importance of feature types. Furthermore, we reviewed the most erroneous predictions (such as ICU admissions with predicted probability of admission less than 0.2) for the primary outcome.

Results:
Our cohort consisted of 41,654 distinct admissions, of which 28,985 (70% of the cohort) were unique patients. In the holdout test set of 10,096 admissions, 75% of patients were not in the training and validation sets. We identified 5,568 (13.4%) admissions who were admitted to the ICUs at some point within 24 hours following their initial admission, and 3,892 (9.3%) in the ICUs at the 24th hour after admission.

Gradient-boosted trees consistently performed better, or just as well as, other algorithm types. Results are summarized in the below table and figure. The full-feature model outperformed the simple-feature model and the ESI-only model. For the
primary outcome, using a threshold such as 0.6, our best model achieved 85% positive predictive value, 89% negative predictive value, 30% sensitivity, and 99% specificity.

<table>
<thead>
<tr>
<th>Models</th>
<th>Primary outcome (highest care level within 24 hours)</th>
<th>Secondary outcome (care level at the 24th hour)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AUROC (95% CI) AUPRC (95% CI)</td>
<td>AUROC (95% CI) AUPRC (95% CI)</td>
</tr>
<tr>
<td>Full-feature</td>
<td>0.88 (0.87 - 0.89) 0.65 (0.63 - 0.68)</td>
<td>0.86 (0.85 - 0.87) 0.50 (0.47 - 0.53)</td>
</tr>
<tr>
<td>Simpler-feature</td>
<td>0.82 (0.80 - 0.83) 0.52 (0.50 - 0.56)</td>
<td>0.81 (0.79 - 0.82) 0.41 (0.38 - 0.45)</td>
</tr>
<tr>
<td>ESI-only regression</td>
<td>0.67 (0.65 - 0.70) 0.37 (0.35 - 0.40)</td>
<td>0.67 (0.65 - 0.70) 0.28 (0.26 - 0.31)</td>
</tr>
</tbody>
</table>

Feature ablation analysis revealed that vital signs were the only stand-alone feature type that significantly impacted the predictive performance of our models. Among patients whose first inpatient admitting physicians were not critical care services, the most erroneous predicted values were for patients from neurosurgery, pulmonary hypertension, and trauma services, accounting for 10 to 18% of all errors.

**Discussion and Conclusion:**

Our model appeared to accurately discriminate patients requiring ICU care against those who did not, which could help identify higher-risk patients and improve clinical decision-making. While the Regime I model outperformed that trained under Regime II (AUROC 0.88 vs 0.82 for primary outcome), the relative ease of feature availability in Regime II makes it an attractive option for robust prediction without complicated data pipelines. Analysis of cases of prediction error indicated greater uncertainty amongst patients initially admitted to certain specialty services, likely due to admission protocols which supersede clinical decision making based on specific disease severity measures.

**Clinical implementation:** We propose a 3-tier alarm system to identify: very low, very high, and uncertain acuity levels depending on case mix and hospital bed capacity. Mid-range patients are those with predicted probability is in the mid-range and would require further assessment and interventions for admission triage.

**Limitations:** Guidelines for admission changes and thresholds for level of care assignments depend on case mix and overall patient acuity. Our models cannot be applied as-is but need to be adapted to specific hospital settings for implementation. Our study is observational, using electronic health record data which has issues such as missing data, entry errors, and inconsistent coding practice.

In conclusion, our proposed prediction models demonstrated good discrimination for ICU vs non-ICU level of care within 24 hours and at the 24th hour after initial admission, substantially better than existing standards of care in patient risk assessment such as the Emergency Severity Index system. Outputs of such models could be used as priority scores to aid triage decision-making, improving the efficiency and quality of emergency hospital care.

**Attendees’ Take-away Tools:**

Our approach and algorithm using electronic health record data for machine learning prediction and proposed three-tier alarm system can be reproduced, adapted, and may be implemented in another practice setting.

**References:**

Developing an Automated Recipient Risk Prediction Tool for Life-long Mortality After Pediatric Heart Transplantation

Yizhao Ni, PhD, Farhan Zafar, MD, Alia Dani, PhD, David Morales, MD
Cincinnati Children’s Hospital Medical Center, Cincinnati, OH

Introduction
Pediatric heart transplantation is a life-saving procedure for children with end-stage heart failure. The procedure has a high early mortality (18%) and the rate remains constant (4% per year) for the next 20 years. The existing studies have focused on models predicting short-term (e.g., 1-year) survival with pre-operative and sometimes operative data. We aim at using state-of-the-art machine learning technologies to analyze patients’ life-long mortality with continual data feeds. Our objective is to develop an automated approach to identify risk factors continually with time period data and predict life-long mortality for pediatric recipients of heart transplantation.

Methods
We utilized heart transplantation data collected by the Organ Procurement and Transplantation Network (OPTN) as of March 1st, 2020. The study was performed in accordance with the regulations of the institutional review board of Cincinnati Children’s Hospital Medical Center. The study population included all pediatric patients aging ≤18 at the time of transplantation. Patient mortality outcome (alive/deceased) was calculated at one month, six months, and then at 6-month increment till ten years to simulate their status during postoperative follow ups. If an alive patient’s record was less than ten years in the dataset, he/she would be excluded from outcome prediction one year after his/her last record. To capture potential risk factors, we extracted patients’ preoperative, operative, and follow-up information from the OPTN data. Descriptive statistics of the extracted variables is presented in Table 1. The numerical variables were standardized using z-score normalization. The categorical variable was binarized with a dummy variable to avoid linear dependencies. To deal with missing data, we implemented grand mean and model imputation for numerical variables and unique-value imputation for categorical variables. If a variable was missing for all patients at a time point (e.g., all follow-up variables were not available at the first month), it was removed from the predictive models.

We aimed to predict mortality outcome for a patient at each time point (time of a follow-up visit) using information up to the patient’s last visit. We developed two approaches to analyze patient information. The first approach was a binary-class classification using a snapshot of latest features till a patient’s last visit. Four traditional machine learning classifiers were implemented including logistic regression (LR) with L1 and L2 normalization, linear support vector machines (SVMs), random forests (RFs), and gradient boosting machines (GBMs). The other was a sequence labeling setting that aggregated information till a patient’s last visit to predict his/her current mortality. We implemented a one direction long-short term memory (LSTM) network to aggregate information and make sequence predictions.

We performed a stratified random sampling based on individual patients to split the dataset into two parts, 70% for training and 30% for evaluation. Ten-fold cross-validation was utilized in training the machine learning and deep learning algorithms, where model parameters were optimized with grid search parameterization. For model evaluation, we used the area under the receiver operating characteristic curve (AUC) as the primary measure. To understand risk factors contributing to patient mortality, we performed an iterative step-forward feature selection (LR with L2 normalization) with best first search to identify key predictors at each time point.

Results
The distribution of patient mortality outcome over time is visualized in Figure 1. The sample size ranged from 10,213 at one-month with a mortality rate of 10% to 6,004 at ten-year with a mortality rate of 48%. The sample size decreased due to alive patients excluded from analysis as described above. Performance of the predictive models is presented in Figure 2. All machine learning models achieved similar performances, and the differences in performance were not statistically significant. The best-performing model (GBM) achieved 0.802 AUC at one-month and over 0.9 AUCs after one year. Figure 3 presents the importance of selected variables over time based on their reciprocal of ranks.

Discussion and Conclusion
The machine learning models showed good capacity for predicting risk of life-long mortality. The best-performing GBM achieved decent AUCs in predicting early mortality and the performance improved over time when more information was available. The LSTM model did not show improved performance, suggesting that mortality outcome was dependent more on recipients’ short-term status. The findings from feature selection suggested shift of importance in risk factors over time. Recipients’ clinical status during transplantation such as ECMO use, stroke and serum total bilirubin were important for early mortality. In the medium and long terms, recipients’ chronic conditions (e.g., renal dysfunction), healthcare utilization and quality (e.g., follow-up care by transplant center, specialty physician, primary care provider, or unknown) and organ matching factors (e.g., recipient age/height, donor weight) were more dominant.
By analyzing time period data, the developed approach provides continual updates of mortality outcome and risk factors. The approach has potential to assist with postoperative care after heart transplantation by informing healthcare providers if a patient requires additional attention, and if so, what are the actionable items.

Table 1. Summary of key variables used in the study.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Subject</th>
<th>Description</th>
<th>Num. Var.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-operative and operative variables</td>
<td>Recipient</td>
<td>Hight, weight, body mass index</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Subject status and time in the waiting list for transplantation</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sociodemographic variables, e.g., age, gender, ethnicity, academic level</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Clinical characteristics such as health conditions, medical histories, lab</td>
<td>111</td>
</tr>
<tr>
<td></td>
<td></td>
<td>results, and functional status</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Donor</td>
<td>Hight, weight, body mass index</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sociodemographic variables such as age, gender, ethnicity</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Clinical characteristics such as health conditions and histories, lab results</td>
<td>77</td>
</tr>
<tr>
<td></td>
<td>Matching</td>
<td>Organ matching assessments between recipient and donor</td>
<td>9</td>
</tr>
<tr>
<td>Follow up variables</td>
<td>Recipient</td>
<td>Hight, weight, body mass index</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sociodemographic variables, e.g., academic level, working for income</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Clinical characteristics and healthcare utilization (e.g., hospitalization)</td>
<td>28</td>
</tr>
</tbody>
</table>

References
Experiences of Discrimination and Withholding Information from Providers
Paige M. Nong, BA¹, Alicia K. Williamson, BA¹, Jody E. Platt, MPH PhD¹, Denise L. Anthony, PhD¹
¹University of Michigan, Ann Arbor, Michigan, USA

Introduction
Patient-centered care depends on provider patient relationships that facilitate information exchange and patient involvement.¹ ² Health information technology (HIT) that promotes the utilization and sharing of patient information, care coordination and quality of care³ ⁴ relies on communication between patients and providers. Patient concerns about what will be done with the information they share, and who has access to it, may inhibit information exchange with providers.⁵ ⁶ For example, patients may withhold information from a provider based on concerns about stigmatized health conditions, security, privacy, or trustworthiness.⁶ ⁷ Incomplete patient data introduces bias, which can negatively impact quality of care.

Some dimensions of trust have been identified as important to disclosure, such as patient trust that a provider will use their personal information appropriately (confidentiality), and will disclose conflicts of interest.⁶ ⁸ Prior work also demonstrates that patients face interpersonal discrimination while seeking medical care⁹ ¹⁰ and that experiences of discrimination are associated with low trust.¹¹ Our study contributes important evidence about the associations between disclosing information to healthcare providers, experiences of discrimination in healthcare, and trust in providers.

Methods
In order to identify the relationships between trust, discrimination, and withholding we fielded a national survey of US adults. A total of 2,157 US adults who speak English from the National Opinion Research Center’s (NORC) AmeriSpeak panel responded to the survey in May 2019 (66% response rate). The final analytic sample included 2,144 respondents and analyses used post-stratification weights based on the Current Population Survey. A more complete description of the survey development and implementation is available elsewhere.⁹

Withholding information was measured as a binary yes/no response to the question “have you ever kept information from your healthcare provider because you were concerned about privacy or security?”. We asked respondents whether they had ever been discriminated against, hassled or made to feel inferior while getting medical care (yes/no). To measure dimensions of trust, respondents indicated their agreement with four statements about providers: 1) “health care providers care most about making money for themselves”, 2) “health care providers disclose their conflicts of interest”, 3) “I trust health care providers to use my health information responsibly”, and 4) “All things considered, health care providers in this country can be trusted”. Responses were coded so that low trust in a given dimension was equal to one and higher trust was equal to zero.

We conducted weighted bivariable and multivariable logistic regressions of withholding information on experiences of discrimination and four dimensions of low trust. Multivariable logistic regressions controlled for demographic, health, and access to care factors: sex, age, race/ethnicity, education, income, self-reported health status, health insurance status, having a regular healthcare provider, high satisfaction with care, and time of most recent visit to a healthcare provider.

Results
In weighted analyses, 27.5% of respondents reported having withheld information from providers because of concerns about privacy or security. Using weighted percentages, approximately half the sample were men (51.8%). The racial and ethnic composition of the sample was 63.8% white (non-Hispanic), 16.8% Hispanic, 11.9% Black (non-Hispanic), 4.7% other and multiple racial/ethnic identities (self-selected), and 3.8% Asian (non-Hispanic). Approximately 19.1% of respondents reported experiencing discrimination while seeking medical care. In weighted bivariable logistic regressions, having experienced discrimination and each dimension of trust are significantly associated with a greater likelihood of withholding information from a provider.

In the multivariable weighted logistic regression, we observe a statistically significant greater likelihood of withholding information from providers for those who report experiencing discrimination in the healthcare system (OR 3.67 [CI 2.6-5.2], p<0.001). Those reporting low trust in providers using health information responsibly (OR
Low trust in providers generally.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Bivariable</th>
<th>Multivariablea</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experienced discrimination in the healthcare system</td>
<td>4.62***</td>
<td>3.67*** [2.62-5.16]</td>
</tr>
<tr>
<td>Low trust in providers’ financial motivations</td>
<td>1.89***</td>
<td>1.13 [0.76-1.70]</td>
</tr>
<tr>
<td>Low trust that providers disclose conflicts of interest</td>
<td>1.94***</td>
<td>1.40* [1.03-1.90]</td>
</tr>
<tr>
<td>Low trust that providers use health information responsibly</td>
<td>3.90***</td>
<td>2.27** [1.41-3.64]</td>
</tr>
<tr>
<td>Low trust in providers generally</td>
<td>2.82***</td>
<td>1.23 [0.74-2.05]</td>
</tr>
</tbody>
</table>

*p<0.001 = ***, P<0.01 = **, P<0.05 = *

a Covariates in the multivariable model include sex, age, race/ethnicity, education, income, self-reported health status, health insurance status, having a regular healthcare provider, high satisfaction with care, and time of most recent visit to a healthcare provider.

**Discussion**

Patients who have experienced discrimination in the healthcare system are significantly more likely to withhold information from their provider. Those with low trust in providers using health information responsibly and disclosing conflicts of interest are also significantly more likely to withhold information. These findings underscore the importance of addressing interpersonal discrimination in the healthcare system and facilitating trustworthy provider relationships in order to ensure patient information sharing necessary for quality patient-centered care.

Incomplete information can result in poorer care, exacerbating existing disparities and compounding the harm of a discriminatory experience. Future work should engage qualitative methods to identify and understand causal relationships and mechanisms, as well as clarify the roles of each type of trust analyzed here. As we improve our understanding of data bias in healthcare and work to minimize it, we must also recognize and address the role of patient trust in providers, confidentiality, and discrimination in healthcare that undermine the quality of care and of healthcare data.

**References**

**PHenotype Observed Entity Baseline Endorsements (PHOEBE)** - recommender system for concept selection in phenotype algorithm development

Anna Ostropolets, MD, MA¹, Patrick Ryan, PhD², George Hripcsak, MD¹,³
¹Columbia University, New York, NY, USA; ²Janssen Research and Development, Titusville, NJ; ³New York-Presbyterian Hospital, New York, NY

**Introduction**

Large-scale observational studies in distributed networks allow leveraging large patient populations to produce reliable evidence, but they require a recognition and reconciliation of differences in patient data capture. While single-center studies can use phenotypic algorithms that were developed and validated on that data source, such algorithms may not be applicable to other data sources. As we have shown before, data sources are highly heterogeneous with respect to ontologies, with most of the concepts being unique to a data source [1]. This challenges conventional approaches to phenotyping [2,3] such as using administrative claims concepts, concepts from existing literature or exploring concepts at a local patient data instance. Due to the large ontological space, patient phenotyping in network studies becomes time-consuming and oftentimes leads to disparate phenotypes representing one clinical idea. Such lack of consistency results in potential patient loss and biased results.

Recent Observational Medical Outcomes Partnership (OMOP) studies [4] enable leveraging a shared knowledge base of concepts allowing the construction of comprehensive phenotypes. We created a recommender system that uses lexical and ontology features to select recommendations and prioritize them based on the prevalence of concepts in the network of 21 US and international data sources. In this work, we will describe its principles and validation experiment.

**Methods**

The details of the Concept Prevalence study that was used to generate the concept knowledge base are described elsewhere [1]. The initial set used by PHOEBE consisted of 11 million unique concepts appearing in at least one source within the network, with 272 billion records summarized to provide verbatim and descendant concept frequency estimates. Aggregated frequencies of concepts across participating data sources were used to pre-compute a set of recommended terms for all standard concepts in the OMOP Standardized Vocabularies. Recommended terms were selected based on string matching and filtered based on the aggregated frequency to remove unused concepts.

First, the recommender system provided the most prevalent starting concept based on the user text input. After the initial concept was selected the OMOP Standardized Vocabulary hierarchy was used to provide ancestor and descendant terms as well as potentially relevant terms not included in the concept set (orphan codes). The output was prioritized based on the aggregated frequency to focus decision-making on the concepts largest gain in number of patients. The recommender system was made publicly available (https://data.ohdsi.org/PHOEBE).

For validation, we used an electronic health record data source (Columbia University Medical Center EHR) and two administrative claims data sources (IBM MarketScan® Multi-State Medicaid Database and IBM MarketScan® Medicare Supplemental Database) translated to OMOP Common Data Model. We selected diabetes mellitus type II and diabetes mellitus type I - two common conditions that were extensively studied in the observational literature and for which specific drug treatment exists. For each of the conditions, PHOEBE was used to create concept sets following the steps outlined above. As a reference, we selected two eMERGE network phenotypes representing the same clinical ideas [5]. For each phenotype, we extracted ICD9-CM concept sets and translated to SNOMED-CT concept [6]. We then created patient cohorts by selecting patients with at least one occurrence of a diagnosis code from corresponding concept sets and with at least 365 days of prior observation to ensure data coverage. For each cohort, we followed patients to look for a specific treatment, which included any occurrence of insulin products for type I diabetes mellitus and oral antidiabetic drugs (metformin, sulfonylureas, thiazolidinediones, dipeptidyl peptidase IV inhibitors and glucagon-like peptide-1 agonists) for type II diabetes mellitus.

We computed the positive predictive value (PPV) of each phenotype, which was defined as a proportion of people with diagnostic code who also had subsequent treatment with corresponding drugs. Additionally, for patients with subsequent treatment identified by both PHOEBE and eMERGE algorithms, disorder start dates were extracted and compared.
Results
We found that for both disorders, the PHOEBE algorithm identified significantly more patients (Table 1).

Table 1. Comparison of eMERGE and PHOEBE concept set performance for diabetes mellitus type I and II. PPV – positive predictive value.

<table>
<thead>
<tr>
<th>Condition</th>
<th>eMERGE cohort</th>
<th>PHOEBE cohort</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Patients with subsequent treatment</td>
<td>Patients without treatment</td>
</tr>
<tr>
<td>Diabetes mellitus type I</td>
<td>CUMC</td>
<td>7,470</td>
</tr>
<tr>
<td></td>
<td>MDCD</td>
<td>178,579</td>
</tr>
<tr>
<td></td>
<td>MDCR</td>
<td>166,600</td>
</tr>
<tr>
<td>Diabetes mellitus type II</td>
<td>CUMC</td>
<td>19424</td>
</tr>
<tr>
<td></td>
<td>MDCD</td>
<td>101,816</td>
</tr>
<tr>
<td></td>
<td>MDCR</td>
<td>105449</td>
</tr>
</tbody>
</table>

While there was no significant difference in PPV for diabetes mellitus type I, PHOEBE algorithm identified more patients with subsequent treatment for diabetes mellitus type II on MDCR and MDCD. As expected, since eMERGE phenotypes were developed on CUMC data, PHOEBE did not show significant patient gain on that data source for both conditions. For patients identified by both algorithms, PHOEBE identified such patients earlier in the course of the disorder on all data sources. In CUMC, patients on average presented with diabetes mellitus type II and diabetes mellitus type I more than two years earlier when using the PHOEBE algorithm (median difference 688 days (IQR 179 – 1,533) and 1532 days (538 – 3,070) respectively). This can be explained by multiple patients having unspecified diabetes mellitus codes prior to being treated by endocrinologists. MDCD and MDCR displayed smaller difference with median difference 374 days (139 - 806) and 407 days (176 - 894) for diabetes mellitus type I and 820 days (377 – 1,431) and 122 days (13 - 417) for diabetes mellitus type II.

One of the limitations of this work is using only US data sources. As PHOEBE leverages data from a large network, it may show better performance on non-US data sources, which is yet to be shown. We used specific treatment as a proxy for patients being true positive. While imperfect, it allows assessing algorithm performance in absence of methods to validate phenotypes in administrative claims data.

Conclusion
Patient phenotyping in network studies in time-consuming and oftentimes produces inconsistent results due to different patient representations in disparate data sources. We developed a recommender system that facilitates phenotype development standardization and comprehensive concept set creation. We examined its performance against eMERGE concept sets and found better performance for some conditions and early detection of patients of interest.

References:
Evaluation of the Portability of Natural Language Processing-based Computable Phenotypes in the eMERGE Network

Jennifer A. Pacheco, MS¹, Luke Rasmussen, MS¹, Ken Wiley², Thomas Nate Person³
David J. Cronkite⁴, Sunghwan Sohn⁵, Shawn Murphy, MD, PhD⁶, Justin H. Gundelach⁵
Vivian Gainer⁶, Victor Castro⁶, Cong Liu⁷, Todd Lingren⁸, Frank Mentch, PhD⁹, Agnes S Sundaresan, MD¹⁰
Garrett Eickelberg¹, Valerie Willis, PhD², Al'ona Furmanchuk, PhD¹
Roshan Patel⁹, David S. Carrell⁴, Marc Williams¹⁰, Elizabeth W. Karlson, MD⁶, Jodell E. Linder, PhD¹¹, Yuan Luo, PhD¹, Chunhua Weng PhD⁷, Wei-Qi Wei, MD PhD¹¹, on behalf of the eMERGE Network Phenotyping Workgroup

¹Northwestern University, Chicago, IL; ²National Human Genome Research Institute, Bethesda, MD; ³Penn State University, University Park, PA; ⁴Kaiser Permanente Washington Health Research Institute, Seattle, WA; ⁵Mayo Clinic, Rochester, MN; ⁶Mass General Brigham, Boston, MA; ⁷Columbia Univ., New York, NY; ⁸Cincinnati Children's Hospital Medical Center, Cincinnati, OH; ⁹Children's Hospital of Philadelphia, Philadelphia, PA; ¹⁰Geisinger, Danville, PA; ¹¹Vanderbilt University, Nashville, TN

Introduction and Objective

Sophisticated natural language processing (NLP) pipelines, such as cTAKES¹ and MetaMap², and simpler rule-based approaches combining regular expressions (RegEx) and logic have been increasingly leveraged in the electronic health record (EHR)-based phenotyping process³. A one-year pilot study was conducted by the electronic Medical Records and Genomics (eMERGE) Network to test the feasibility of deploying portable phenotype algorithms that are enriched by information from NLP, with the goal of improving the performance of existing algorithms. We report our results including lessons learned, and recommendations of best practices for incorporating portable NLP algorithm components in computable phenotypes.

Materials and Methods

Based on scientific merit and predicted difficulty, eMERGE selected 6 existing phenotypes to enrich with NLP. The eMERGE Phenotyping workgroup restricted NLP pipelines to cTAKES, MetaMap, and/or regular expressions to assist in algorithm portability, and used 2 commonly adopted negation detection modules: NegEx⁴ and ConText⁵. Algorithms were developed, then validated by at least one other site, before dissemination to up to 8 other sites for implementation, similar to our existing validation workflow⁶. We retrospectively compared the NLP methods and tools to assess performance, portability, and ease of use. Sites reported lessons learned for creating and sharing NLP algorithms via a brief informal survey designed to collect both quantitative and qualitative data. Quantitatively, sites reported changes in performance, and estimated the amount of resources and elapsed time it took to complete.

Using grounded theory⁷, a thematic analysis was conducted by two authors (JAP, LVR) via independent review of the qualitative feedback. First, open and axial coding on categories of issues or concerns was completed to identify phrases and loosely categorize them. The coders used selective coding to refine axial codes into a comprehensive hierarchical codebook, independently re-coded the feedback, and reviewed to achieve consensus. Emergent themes were identified through iterative review of the codes. Finally, we summarized lessons learned for review by all authors, including a) challenges for each theme; b) corresponding best practices to address those challenges based on existing published evidence and/or eMERGE experience; and c) opportunities for future research.

Results

In all but one case (where precision decreased), the precision and recall stayed the same or improved. Of note, the time it took to develop and validate phenotype algorithms at secondary sites was considerably longer when NLP was
incorporated (1-2 years). Figure 1 shows the three major themes identified from the qualitative analysis: portability, phenotyping, workflow/process, and technology, along with sub-themes. The technology theme was used as a modifier for the other two primary themes - that is, all technologies were associated with another theme.

Figure 1. Themes of lessons learned

Discussion and Conclusion

Our study shows that an NLP component brings extra benefits to an existing phenotyping algorithm. This study also demonstrates NLP performance may vary between primary and validation sites, due to heterogeneity in clinical document names and the basic structure of clinical notes. Implementing standard terminologies to categorize all document types is impractical due to the absence of clear document types at this stage. Thus, we suggest starting with semi-structured clinical notes and imposing minimal reform or costs at the institutional level, such as comprehensive documentation and customization needed at each site. The generalizability of a negation module remains an open NLP challenge; thus, local tailoring on negation may be necessary, such as adding correction rules to the code for negating language. It is also important to understand that to successfully share and implement a computable phenotype using NLP is not just about the NLP technology or the algorithm itself. Critical factors to ensure project success include privacy protection and attention to regulations; intellectual property agreement; and technical infrastructure setup and efficient communication as previously reported.

Limitations to this study include that we preselected pipelines for each task as the comparison of the performance using other pipelines was beyond our resources and timeline. Finally, no formal, standardized measurement of time and effort was used, leading to reliance on estimates which could lead to inconsistent reporting and inaccuracies.

Our results show that NLP continues to make a positive impact on phenotype performance. While portable phenotype definitions are possible, careful planning and architecture of the algorithms is essential to support local customizations that are expected to be needed for the foreseeable future.

References

Statistical Modeling of Multiple Vital Sign Trajectories to Assess Risk of Postoperative Complications and Predict Readmission

Rema Padman, PhD1; Sameera Kodi, MS1; Urmila Ravichandran, MS2; Nirav Shah, MD, MPH2,3
1Carnegie Mellon University, Pittsburgh, PA; 2NorthShore University HealthSystem, Chicago, IL; 3University of Chicago, Chicago, IL

Introduction
Early identification of patients at risk for postoperative complications can facilitate timely workups and treatments and improve health outcomes. Current risk prediction approaches mostly use only preoperative predictors to assess the risk of postoperative complications. They lack dynamic, real-time capabilities to include postoperative information, such as postoperative vital signs, that are collected as part of routine care during the entire inpatient stay. Studies have shown that postoperative fevers, for example, result in increased diagnostic testing, antibiotic usage and length of stay. These may, in turn, result in decreased patient safety, increased antimicrobial resistance and readmission, higher healthcare utilization, and increased morbidity and mortality. With millions of inpatient surgeries completed per year in the United States, improving clinicians’ ability to target only those patients at highest risk of complications has the potential to improve patient safety, quality of care and value on a large scale. In this study, we apply multi-trajectory modeling and evaluation of elective colectomy patients using multiple postoperative vital signs to analyze post-surgery complications and predict risk of readmission.

Methods
Data were extracted for all patients who underwent elective colectomy between January 1, 2007 and December 31, 2013 at NorthShore University HealthSystem. For each surgical episode, information regarding the index hospitalization and any subsequent patient encounters within 30 post-operative days were extracted from the electronic data warehouse (EDW). These included demographics, comorbidities, vitals, laboratory results, medications, length of surgery, start and end time of surgery, length of hospital stay (LOS), readmissions and death. Complications data was extracted initially from the EDW, and then confirmed by manual chart review by a physician, using criteria of symptoms and objective data from diagnostic testing. Missing data was handled using imputation, when feasible, such as missing body mass index (BMI) using the average for the age group and gender of the patient. Group-based multi-trajectory modeling was applied to study the joint developmental course of patient temperature, mean arterial pressure, pulse and WBC readings during inpatient stay up to 9 days (mean LOS < 15% inpatients after 9 days) and readmission risk was predicted for each group. Analyses include: (1) Trajectories of temperature, mean arterial pressure, pulse and WBC measurements taken every four hours, using a censored normal distribution, with important patient characteristics including age, gender, BMI, the actual surgery procedure time, and comorbidities that were considered potentially relevant for the surgical condition such as diabetes, hypertension, chronic obstructive pulmonary disease, rheumatoid arthritis, congestive heart failure, and metastatic cancer, and (2) Comparison of complication rates by trajectory group (3) Comparison of readmission prediction from the multi-trajectory model against five standard machine learning models, with SMOTE for handling imbalance in the data, and five-fold cross validation to obtain readmission probabilities.

Results
Postoperative vital sign trajectories from 584 patients who underwent elective colectomy were evaluated for risk of postoperative complications. A group-based multi-trajectory model with 5 distinct trajectory groups was generated using temperature, mean arterial pressure, pulse and WBC measurements over a period of 9 days as shown in Figure 1. Number of groups was determined using Bayesian Information Criterion. Groups 1 (18.5%) and 2 (30.6%) include low risk patients whereas group 4 (20.2%) contains medium risk patients. Groups 3 (19.9%) and 5 (10.8%) include high-risk and very high-risk patients, respectively. Table 1 shows the profiling of groups according to patient demographics and comorbidities. We see that high-risk groups 3 and 5 had longer procedure hours and more patients with diabetes and pulmonary disease. Low risk groups 1 and 2 had younger patients, more females and fewer rheumatoid arthritis patients. Group 4 had highest percentage of congestive heart failure patients. Also, the post-operative fever percentages are higher in the higher risk groups who also experience more readmissions. Table 2 summarizes the complications, showing a general trend in increasing complications with increasing risk. In particular, we note that deep vein thrombosis, surgical site infection and bacteremia are more prevalent in high-risk groups. For readmission prediction, we observe the predicted probabilities from the prognostication model increasing from low to high-risk groups in Table 3. The group-based prognostication model also exhibits a better AUC score in Table 4 compared to the standard machine learning models evaluated in this study.
The multi-trajectory model stratifies colectomy patients into distinct risk groups, based on a combination of postoperative vital signs and baseline covariates, that are correlated with multiple postoperative complications. The prediction model for readmission risk, in particular, modestly outperforms some widely used machine learning models and provides a more interpretable solution with better insights. These results demonstrate significant promise for developing a more accurate, robust and interpretable postoperative complication and readmission risk approach which can be deployed, similar to NSQIP, to benefit the millions of annual surgeries in the US to potentially lower adverse events and healthcare costs and improve surgical outcomes. This study is limited by its retrospective nature and the inherent limitations of obtaining data from an EDW.

Acknowledgements
We thank Heinz College graduate students S. Yazdi, P. Prosapio, A. Sahoo, Y. Xu, and P. Raja for their early assistance with the exploratory analysis of the post-surgery colectomy data.

References
Addressing Disparities in Diabetes Using Temporal Fairness Models

Joseph M. Plasek, PhD1, Chunlei Tang, PhD1,2, Xiaoxia Wang1, Yun Xiong, PhD3, Yangyong Zhu, PhD3, Yanning He, MD1, Patricia C. Dykes, PhD, RN, FAAN, FACMI1, David W. Bates, MD, MSc1,2, Li Zhou, MD, PhD1

1Division of General Internal Medicine and Primary Care, Brigham and Women’s Hospital, Harvard Medical School, Boston, MA, USA; 2Clinical and Quality Analysis, Mass General Brigham, Boston, MA, USA; 3Shanghai Key Laboratory of Data Science, School of Computer Science, Fudan University, Shanghai, China; 4Department of Endocrinology, Yueyang Hospital of Integrated Traditional Chinese Medicine and Western Medicine, Shanghai, China.

Introduction

Diabetes is an important global public health burden. In the U.S., 10.5% of the population or 34.2 million individuals have diabetes, and 34.5% of the adults or 88 million people aged 18 years or older have prediabetes1. Health disparities in diabetes and its complications (or/and co-morbidities) exist worldwide2-3. Biological, behavioral, social, environmental, and health system contributors (e.g., social determinants of health) to diabetes disparities rely on sensitivity and specificity and are subject to clinical intervention and insurance product development. With the adoption of machine learning in healthcare, the potential to increase knowledge of this condition illuminates the differences in how diabetes presents, its variable prevalence, and how best treatment practices vary between subpopulations4. Simply training a learning model to minimize prediction error has been criticized for its potential to exacerbate unintended discrimination based on sensitive characteristics. A cause of criticism is that machine learning applications often consider static objectives defined on a snapshot of the population at one instant in time, resulting in static data bias being captured5. Temporal fairness models via fairness criteria evaluate the trend of long-term outcomes and impose selected fairness constraints. In this talk, we discuss how temporal fairness models might reduce the disparities in diabetes care or at least avoid perpetuating discriminatory practices.

Methods

We retrieved 420,979 unique patients with diabetes seen between July 1st, 2017, and June 30th, 2018, gathered from the research patient data registry at Mass General Brigham (MGB), a large integrated healthcare delivery network based in Boston, MA. Our dataset’s population characteristics show that most patients are Caucasian (69.5%) but that sizeable subpopulations of minorities (i.e., African American (5.72%), Hispanic (5.293%), Asian (3.03%), Other/Not Recorded (16.51%)) exist. We extracted patient demographics, insurance plan data, and diagnostic information corresponding to the fields of [principal diagnosis], [admitting diagnosis], and [diagnosis 1...n] that are coded as International Classification of Disease (ICD10) codes. This study was approved by the MGB IRB.

We used a three-step, pipeline approach to: (1) conduct a hypothetical diabetes scenario to approximate the intended relation between input and output values; (2) extract the outcome curves using an optimized version of Liu et al.’s delay model6; and (3) analyze each fairness criteria (and their possible optimizations) over time. Fairness criteria include utility-maximizing criterion (MU), demographic parity (DP), equal opportunity (EO). MU ensures the institution is free to focus solely on utility. DP results in equal selection rates between both groups. EO gives the conditional probability of selection to patients who can be successfully independent of the population, resulting in equal true positive rates between both groups. We assume that the institution is utility-maximizing but may impose certain constraints to ensure that the decision is fair. In many cases, it may be unrealistic for an institution to ensure that fairness constraints are met exactly. Liu et al.’s model is a one-step feedback model that considers temporal indicators; it was first introduced in a hypothetical lending scenario. When we experimented with Liu et al.’s model, we found that it required tuning of too many parameters, and thus we could not apply it directly to our diabetes scenario. We optimized the input layer of Liu et al.’s model to only utilize the source, success, buff, and utility parameters. We consider a total population that is comprised of two groups A and B, and an institution (i.e., hospital) that makes a binary decision for each diabetes patient in each group (the source parameter). Diabetes patients in each group are assigned scores and the score for this group follows a type of probability distribution. The success parameter is defined as any diabetes patient with a diabetes-related ICD10 code (“E1”). The buff parameter measures whether a patient is fortunate or not, associated with success and could be affected by multiple factors such as racial or ethnic difference.
in selecting an insurance product. The utility parameter represents decreased disease management expenses. The model output is a visualization depicting an outcome curve showing the average score change.

Results

Figure 1 displays outcome curves corresponding to four ethnic groups and the three fairness criteria. This output can be interpreted, for example, in MU, when enough diabetes patients in a group are successfully given an accurate diagnosis and appropriate management plan, the average score in that group is likely to decrease (MU results in a negative average score change in the population in this case); as we deviate from profit maximization to give out diagnoses to more patients, the average score change decreases up to a certain point, where it is minimized (the altruistic optimum). The MU is primarily in a state of relative harm, consistent with our assumption of “no active harm.” DP and EO are much better (both residing in a state of “relative improvement”). Interestingly, the curves for African American and Hispanic are similar to each other, but different from the Asian and Caucasian curves (which also have a similar topology).

Figure 1 Disparities across ethnic groups when placed under fairness criteria constraints (compared with a schematic diagram of the outcome curve). Note that the horizontal axis represents the selection rate for patients related to the distributions of both score and fairness criterion; the vertical axis represents the mean change in source.

Discussion

We present an adaptation of an innovation from the economics domain, an optimized version of Liu et al.’s temporal fairness model, to simulate and understand how the fairness of algorithms might mitigate or perpetuate undesirable societal impacts in healthcare. Our main finding is that disparities exist as MU results in relative harm but that this might be mitigated with an unconstrained utility objective (DP & EO). A temporally dynamic model fairness paradigm should be adopted to assess how fair the model is when treating pre-existing biases in the data distribution while simultaneously accounting for the impact decisions have on the underlying population over time. This solution might help to mitigate malpractice risk, improve insurance product offerings, or optimize population health management.

Acknowledgments: We thank Sharmitha Yerneni and Carlos Ortega for help drafting the IRB protocol.

References

Consequences of Rapid Telehealth Expansion in US Nursing Homes: Exploring Stakeholder Perspectives

Kimberly R. Powell, PhD, RN1 & Gregory L. Alexander, PhD, RN, FAAN, FACMI, FIAHSI2
1Sinclair School of Nursing, University of Missouri, Columbia, MO; 2Columbia University School of Nursing, New York, NY

Introduction: The COVID-19 pandemic has thrust telehealth, the provision of health care resources, tools, and consultation via mobile digital technologies1, from a convenience to an imperative. Before the COVID-19 crisis, telehealth uptake in nursing homes (NH) was slow and understudied.2-4 While expectations regarding use and potential benefits of telehealth in NH is high, experience shows unplanned and unexpected consequences can occur as a result of major policy and technological changes.5 In March 2020, Medicare rules and regulations were relaxed to broaden access to telehealth services in NH settings so that beneficiaries receive the care they need virtually, avoiding risky travel and possible exposure.6 Augmented clinical work resulting from telehealth expansion during the COVID-19 pandemic warrants urgent and deliberate investigation exploring the experience of NH staff and providers interfacing with these mobile systems. Using the AMIA Input-Output model of unintended consequences5 as our theoretical framework, our aim was to describe perceived anticipated and unanticipated consequences of rapid telehealth expansion in NH.

Methods: Using a mixed sequential explanatory design, we drew a sample based on findings from an analysis of telehealth expansion in a national study examining trends in NH IT maturity.7 IT maturity describes IT capabilities, use and degree of integration in a NH setting. We used maximum variation sampling to purposively select participants who: 1) were affiliated with a NH who participated in our IT maturity survey for two consecutive years 2) completed year 1 of the IT maturity survey prior to telehealth expansion (before March 6, 2020) and year 2 after telehealth expansion (after March 6, 2020) 3) represented a broad range of facility characteristics (e.g., location, bed size) 4) were identified as an end-user of telehealth or responsible for telehealth implementation at the NH. Using six questions from the IT maturity survey, we created a total telehealth score (THS) for each facility. We selected participants from NHs with change in THS ranging from year 1 -year 2 in the bottom 20% (low THS), middle 60% (middle THS), and upper 20% (high THS) to enhance variability. An interview guide was used to ensure consistency among participants and all interviews were conducted using Zoom video conferencing. Each interview was transcribed verbatim by members of the research team with prior transcription experience. A second reviewer verified the transcripts with the audio recording to ensure accuracy. Data were analyzed using directed content analysis. The coding process began by highlighting exact words from the text that appeared to capture key thoughts or concepts. The preliminary codebook was iteratively refined as additional transcripts were analyzed. Preliminary codes were sorted into categories by identifying relationships between codes and groupings of codes. Dedoose qualitative software was used to help organize and visualize data. Processes to ensure trustworthiness were maintained by ensuring credibility, dependability, confirmability, and transferability.8 This study was approved by the University of Missouri institutional review board project number 2009109.

Results: Interviews were conducted with (n=20) NH administrators and clinicians from 16 facilities. Participant demographics are reported in Table 1. Participants represented NHs with change in THS ranging from -22 to +17 (median 3) from year 1 to year 2. Over the course of the analysis, 42 open codes were developed. These were grouped by theme and organized according to the domains of the Input-Output model (See Figure 1). Desirable/anticipated consequences included benefits of avoiding travel for NH residents (e.g., avoiding exposure to infectious disease, avoiding stress associated with leaving the NH, residents more comfortable with NH staff) and saving NH resources (e.g., not sending staff out, avoiding transportation). Desirable/unanticipated consequences included improved access to care and enhanced communication. Undesirable/unanticipated consequences were worsening social isolation, difficulty for residents with cognitive impairments and workflow/technology usability challenges. We found similarities and differences in perceptions of TH expansion by participants from NHs with low, middle, and high THS. Highlighted cells (see Figure 1) indicate the top code(s) for each theme according to THS group.

Discussion: Identifying unintended consequences of rapid TH implementation in NHs is a necessary prerequisite for developing strategies to mitigate undesirable consequences and providing data needed for developing best practices
based on desirable consequences. Mitigation strategies for undesirable consequences could include better standardization, use of secure data exchange, workforce development, identifying types of visits (initial vs. follow up) and types of residents best suited for TH, workflow (using NH staff vs. provider medical assistant in the facility) and measures to ensure resident privacy. Identifying unanticipated/desirable consequences, such as opportunities for early intervention and improved communication should be considered when evaluating factors contributing to the success of TH use in NHs. The qualitative methods used in this study are well-suited to studying the complexity and contingencies that emerge when technological tools, like TH, are rapidly implemented in real-world environments. Further research is needed into the unintended consequences of TH implementation in NH and the widespread dissemination of these findings is even more pressing now as federal initiatives to enhance safety and care requirements incentives, like those set forth in the Nursing Home Reform Modernization Act of 2020, are being considered.

Figure 1. Themes and categories with Exemplar Quotes

<table>
<thead>
<tr>
<th>Main Theme</th>
<th>Sub-Theme</th>
<th>Category</th>
<th>Exemplar</th>
<th>Total Telehealth Score (THS) Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Desirable/ Anticipated</td>
<td>TH saves resources</td>
<td>Not sending staff out</td>
<td>&quot;In some respects, then it helps us with staffing because we always send somebody, an employee with them when they go out and we don't have to do that. So that helps us from a staffing standpoint.&quot; (HR, administrator)</td>
<td>Low (n=12) Middle (n=22) High (n=6) Total</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Avoiding transportation</td>
<td>&quot;The biggest advantage is our transportation costs have significantly decreased because we're not doing this, just doing so much for transportation. That was the big issue.&quot; (HR, administrator)</td>
<td>22.3% 61.8% 15.9% 100%</td>
</tr>
<tr>
<td>Desirable/ Unanticipated</td>
<td>Improved access to care</td>
<td>Greater access to providers</td>
<td>&quot;With mental health, it's meant more timely appointments, it's been more responsive. Prior, we'd have quarterly appointments and this way if she has a few minutes we can squeeze in an evaluation as needed.&quot; (HR, administrator)</td>
<td>9.8% 27.2% 63.0% 100%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Opportunity for timely</td>
<td>&quot;If there's a change of condition, I think it's great because you've got a little more access quickly to the physician than you do have to wait for an appointment or waiting for someone to come to the phone, or whatever. I think it is better.&quot; (NH administrator)</td>
<td>36.7% 50.5% 13.8% 100%</td>
</tr>
<tr>
<td>Undesirable/ Unanticipated</td>
<td>Workflow and tech</td>
<td>Lack of available</td>
<td>&quot;For a lot of doctors, they have to end up charting in two systems, so we're charting in the hospital, health systems EHR, and accessing the nursing home's EHR. I have to literally go into the EHR of the nursing home, pull up all the data and literally write it over yourself, the vital signs, I look and see what nursing said.&quot; (CC, clinician)</td>
<td>10.0% 60.0% 30.0% 100%</td>
</tr>
<tr>
<td></td>
<td>usability challenges</td>
<td>integrated data</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Lack of TH training</td>
<td>&quot;I haven't really seen that many telehealth training courses, or how do you do the best interviews. We've been thinking about that, who should we pull to teach us.&quot; (CC, clinician)</td>
<td>46.2% 42.5% 11.3% 100%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Difficulty to schedule</td>
<td>&quot;Now they'll call and say, &quot;Hey, can I do this in 10 minutes? Can somebody do this right now?&quot; If they're looking at something and they want to see somebody, they're just like, &quot;I can do that right now.&quot; Well, they can, but it just means our nurses are like, oh, I might be already in the middle of doing one or, can you call back in a little bit to do this, so trying to reorganize structured scheduling has been a little bit more complicated.&quot; (TH administrator)</td>
<td>20.0% 60.0% 20.0% 100%</td>
</tr>
</tbody>
</table>

Table 1. Participant Demographics (n=20)

<table>
<thead>
<tr>
<th>Participant title</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Administrator</td>
<td>14 (70%)</td>
</tr>
<tr>
<td>Clinician (nurse or physician)</td>
<td>6 (30%)</td>
</tr>
<tr>
<td>Number of years in current position: range (median)</td>
<td>1.5-24 years (3.5)</td>
</tr>
<tr>
<td>NH facility characteristics</td>
<td></td>
</tr>
<tr>
<td>Change in telehealth score</td>
<td></td>
</tr>
<tr>
<td>(Y1-Y2)</td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>4 (20%)</td>
</tr>
<tr>
<td>Middle</td>
<td>10 (50%)</td>
</tr>
<tr>
<td>High</td>
<td>6 (30%)</td>
</tr>
<tr>
<td>Bed size</td>
<td></td>
</tr>
<tr>
<td>Small (&lt;60)</td>
<td>4 (20%)</td>
</tr>
<tr>
<td>Medium (60-120)</td>
<td>10 (50%)</td>
</tr>
<tr>
<td>Large (&gt;120)</td>
<td>6 (30%)</td>
</tr>
<tr>
<td>Location</td>
<td></td>
</tr>
<tr>
<td>Metro</td>
<td>14 (70%)</td>
</tr>
<tr>
<td>Rural</td>
<td>6 (30%)</td>
</tr>
</tbody>
</table>

References
Among service members transitioning from military to civilian life, two adverse outcomes indicative of poor reintegration are homelessness and criminal justice involvement. Veterans have historically experienced homelessness at a higher rate compared to non-veterans \(^1\) though the rate of veteran homelessness has dropped substantially in the past five years in large part due to interventions by the Department of Veterans Affairs (VA).\(^2,3\) Risk factors for homelessness in veterans include posttraumatic stress disorder (PTSD),\(^4\) severe mental illness (SMI; psychotic, bipolar, and major depressive disorders), misconduct, substance abuse, and traumatic brain injury (TBI).\(^2\) With respect to criminal justice involvement (CJI), risk factors in veterans include younger age, male sex,\(^5\) military experience,\(^6,7\) substance abuse,\(^5,6\) PTSD,\(^6,7\) and SMI.\(^8\) While existing programs within VA can use health system data to help identify individuals at risk of adverse outcomes this is not possible for those transitioning from the Department of Defense (DoD) because linked data are not currently available.

To address this we conducted a longitudinal cohort study that used data from the DoD health system to identify risk factors for homelessness and criminal justice involvement within 2 years of separation from the military for Post-9/11 era Veterans in VA care. Variables from DoD data included demographics (age, sex, race/ethnicity, marital status), military characteristics (branch of service, highest rank and indicator for deployment to a combat zone) and health status documented in DoD care using ICD-9/ICD-10 codes (TBI, substance use disorder (SUD), PTSD and other mental and physical health conditions). Homelessness and criminal justice involvement were identified in VA health system data using ICD-9/ICD-10 codes and VA clinic stop codes for homelessness and criminal justice clinics. To examine the association of neighborhood deprivation of initial post-discharge location we included the Area Deprivation Index (ADI) from the Neighborhood Atlas developed by the University of Wisconsin. We used Fine and Gray competing events models to identify characteristics associated with time to homelessness or criminal justice involvement within two years of military discharge. We present data below for variables with an adjusted hazard ratio \( \geq 1.5 \) or \( \leq 0.67 \).

Among the 418,624 veterans who met inclusion criteria, the two-year incidence of homelessness and criminal justice involvement was 4.7% and 0.8%, respectively. Homelessness was most strongly related to being Black (2.24 [2.16-2.31]) or Native American/Pacific Islander (1.54 [1.51-1.61]), being unmarried (1.56 [1.51-1.61]), being 30 or younger (vs. 40-49 (0.43 [0.40-0.46]) or \( \geq 50 \) years (0.26 [0.21-0.32]), being enlisted (2.38 [2.13-2.63]), SUD (2.08 [2.01-2.14]), severe mental illness (SMI 1.93 [1.85-2.00]), and personality disorder (1.71 [1.60-1.82]). Criminal justice involvement was most strongly associated with male sex (2.94 [2.50-3.45]),
being unmarried (1.84 [1.70-1.99]), being 30 or younger (vs. 40-49 (0.12 [0.04-0.32]) or ≥50 years (0.34 [0.29-0.40]), being enlisted (1.96 [1.52-2.56]), SUD (3.53 [3.27-3.81]), SMI (1.96 [1.80-2.15]), TBI (1.58 [1.40-1.78]), and military deployment (1.85 [1.66-2.06]). Veterans with co-occurring SUD, SMI, and posttraumatic stress disorder had 4 times the incidence of homelessness and 10 times the incidence of criminal justice involvement vs. veterans with none of these diagnoses. Moreover, while the impact of the ADI was not significant by itself, the incidence of homelessness and criminal justice involvement was highest among veterans with co-occurring mental health and SUDs residing in disadvantaged neighborhoods.

These findings identify the importance of using informatics tools to assist in identifying risk factors for adverse outcomes for service members as they transition to civilian life. Programs beginning before transition are needed to provide additional, targeted assistance to Black and Native American/Pacific Islander Veterans and those with complex mental health comorbidity. Further research may be warranted to better understand how existing or new programs/policies may help address needs of Veterans with combat deployment, TBI and complex mental health comorbidity regarding criminal justice involvement.

References:

Acknowledgements: This study was funded by DoD CDMRP (W81XWH-18-1-0247). Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the U.S. Government, the Department of Defense or the U.S. Department of Veterans Affairs, and no official endorsement should be inferred.
Planning Equitable Clinical Trial Enrollment using Integer Programming

Miao Qi, BS1, Amar K. Das, MD, PhD2, Kristin P. Bennett, PhD1
1Rensselaer Polytechnic Institute, Troy, New York, USA; 2Center for Computational Health, IBM Research, Cambridge, Massachusetts, USA

Introduction

Ensuring equity in randomized clinical trial (RCT) participation is an important societal goal. Prior informatics work has focused on identifying potential underrepresentation of minorities and other groups in RCT recruitment. For example, in GIST 2.01, the rates of subgroups in the target population and those meeting eligibility criteria are estimated from electronic health records (EHRs), and the difference between actual and target rates may reveal potential inequities. We propose an alternative approach assessing target subgroup rates that is focused instead on designing and achieving RCT enrollment plans that meet equity targets. We show that we can derive the equity targets for multiple subgroups specified over the superset of protected attributes in a patient population. These attributes can include demographic, clinical, laboratory, or risk factors. We use an integer programming (IP) model, integrated with the equity targets and total recruitment goal, to create an equitable enrollment plan. For this presentation, we discuss our approach and demonstrate its application to a previously conducted RCT study of hypertension.

Methods

Given data from clinical patient population (e.g., a health system’s EHR) or a nationally representative population of disease prevalence (e.g., one based on the National Health and Nutrition Examination Survey, or NHANES), we have developed an approach to define rates for all possible subgroups in disease-specific set (S). We use these rates to produce an equitable enrollment plan for the desired number of participants (n_j) across all subgroups j ∈ S. Specifically, we formulate an IP model that informs, for a given trial size, whether it is feasible to meet target enrollments (n_j,ideal) estimated from the patient population, and, if not, creates an enrollment plan that is equitable as maximally possible. The IP model takes into account desired constraints in the RCT design such as minimum subgroup sizes (n_j,min). The model can also determine trial sizes that are likely to meet equity targets for all subgroups of interest.

\[
\begin{align*}
\min_{N,z,d} & \quad \sum_{j \in \hat{S}} (\delta_j z_j + \lambda_j d_j) \\
\text{s.t.} & \quad l_j - z_j \leq n_j \leq u_j + z_j \\
& \quad n_j,ideal - n_j \leq d_j \\
& \quad n_j - n_j,ideal \leq d_j \\
& \quad \sum_{j} n_j \leq N \\
& \quad z_j, d_j \geq 0 \text{ are integer} \\
\forall j \in \hat{S}, & \quad n_j \geq n_j,\text{min} \geq 0 \text{ are integer} \\
\forall j \in \hat{S} \setminus S, & \quad n_j = \sum_{k \leq j} (n_k)
\end{align*}
\]

The IP minimizes the difference between RCT and target subgroup distributions using at most N participants. Let x be a set of patient attributes, S be the set of all lowest-level subgroups defined over x (i.e. every attribute takes unique value) and \( \hat{S} \) be the set of all possible non-empty subgroups defined over x. For each subgroup \( j \in \hat{S} \), we use 95% CI based on equity measures to define an equity target range \( (l_j, u_j)^2 \). Subgroup \( j \)'s enrollment size, outside of its range, generates the equity violation \( z_j \).

The absolute difference between the plan and real RCT subgroup size, \( d_j \), ensures the model does not stop at boundary values when better choices are available. The first model constraint computes the equity violation to be minimized; the next two constraints make sure the subgroup size is as close to the target size as possible; then, we assure enrollment numbers are reasonable in real world; finally, the summation of subgroups are required to equal its parent group. Since patient subgroups are hierarchical based on the attributes, computation of lowest-level subgroup assignments is sufficient; higher-level subgroup assignments can be derived from their children. Penalty coefficients \( \delta_j \) and \( \lambda_j \) can be adjusted based on different needs. For example, if overrepresentation is considered less harmful than underrepresentation, we set penalty coefficients for underrepresented groups larger. These coefficients provide flexibility to model to real-world requirements.

To demonstrate our approach, we calculate survey-weighted prevalence rate estimates for each subgroup within the U.S. population of individuals with hypertension using NHANES 2015-2016\(^3\). These estimates form equity targets, which could alternatively be derived from a local clinical population or a global population. We assess our IP model...
results against the participant-level data from the Systolic Blood Pressure Intervention Trial (SPRINT)\(^4\), available from the Biologic Specimen and Data Repositories Information Coordinating Center (BioLINCC). The difference between subgroup equity targets and RCT enrollment can be measured using metrics\(^{1,2,5}\) such as GIST 2.0 (difference in rates of RCT and target), or a machine learning-based fairness metric (difference in log odds of RCT versus those of target). We have implemented the latter metric in an R Shiny-based tool. For categorization and visualization purposes, our tool shows for any planned, ongoing, or completed trial the absence (red), high underrepresentation (orange), underrepresentation (light orange), equitable (teal), overrepresentation (light blue), and high overrepresentation (blue) based on a user-defined metric with lower and upper thresholds \((\tau_l,\tau_u)\) and associated statistical test\(^2\).

**Results**

The following table shows enrollment targets for subgroups using the same trial size \(N = 9361\) as the original SPRINT study with \(x = \{\text{race/ethnicity, age}\}\). The table indicates the actual trial enrollment and the equity measure for each subgroup in the SPRINT trial and an enrollment target and range that would be equitable. Equity = subgroup \(\log((\text{odds in trial})/(\text{odds in equitable plan})).\(^1\)

<table>
<thead>
<tr>
<th>Subject Characteristics</th>
<th>SPRINT Trial</th>
<th>Equitable Enrollment Plan</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Enrollment</td>
<td>Equity</td>
</tr>
<tr>
<td>All 18-39</td>
<td>0</td>
<td>Missing</td>
</tr>
<tr>
<td>Hispanic 18-39</td>
<td>0</td>
<td>Missing</td>
</tr>
<tr>
<td>Hispanic 40-59</td>
<td>299</td>
<td>-0.606</td>
</tr>
<tr>
<td>Hispanic 59+</td>
<td>686</td>
<td>0.355</td>
</tr>
<tr>
<td>Hispanic All</td>
<td>985</td>
<td>-0.045</td>
</tr>
<tr>
<td>NH Asian All</td>
<td>79</td>
<td>-1.600</td>
</tr>
</tbody>
</table>

**Discussion**

Our approach helps clinical researchers incorporate equity considerations into enrollment plans by identifying optimal numbers of participants in each subgroup for a given study size. The equity targets can be used to specify enrollment of race/ethnicity and gender subgroups for the NIH-required Planned Enrollment form. Our tool allows researchers to review and visualize how close the equity of an ongoing RCT recruitment effort is to the target. The tool is adaptable to meet other trial-specific goals. Future work includes developing a variant of the IP model to address multi-site trial planning, in which the number of participants that are enrolled at each site should meet overall equity targets.

**References**


\(^1\) Settings: \(n_j, n_{min} = 5, \delta, \chi = 1, \tau_l = 0.2, \tau_u = 0.4. \) NH = Non-Hispanic.
COVID19 Hospital Length of Stay, Readmission Rate, and Hospital Death Tied to Socioeconomic Area of Deprivation Index

Priya Ramaswamy, MD MEng1, Phoebe Lu, MPH1, Jen J Gong, MS PhD1, Travis Porco, PhD MPH1, Priya B. Shete, MD MPH1, Seth Blumberg, MD PhD1, Xinran Liu, MD MS FAMIA1
1University of California, San Francisco, California, USA

Abstract

Higher Area of Deprivation Index (ADI), a socioeconomic measure based on US census data, is often associated with worse health outcomes. For 822 COVID-19-positive inpatient encounters at a California hospital system, we found evidence of relationships between ADI and 30-day readmission (p<0.001), in-hospital death (p<0.001), and hospital length of stay (p<0.001). Thus, disparities in the acute care of COVID patients should be addressed.

Introduction

There has been increased national interest in reducing healthcare disparities based on socioeconomic background, especially with the onset of the COVID-19 pandemic which has disproportionately impacted minorities and poorer neighborhoods.1

Area of Deprivation Index (ADI), a socioeconomic measure whose granularity is at the neighborhood (9-digit zip code) level, is derived from US census data aggregating factors like education, employment, income, poverty, and housing. There have been several studies focusing on aggregate COVID-19 positivity rates based on ADI, but less attention has been given to in-hospital outcomes. If such outcome disparities exist at the inpatient encounter level, then measures should be taken by health systems to improve COVID-health equity. We address this gap by measuring the relationship between Area of Deprivation Index (ADI) and inpatient outcomes such as hospital length-of-stay (LOS), all-cause readmission rate, and in-hospital death at a tertiary academic center.

Methods

After obtaining IRB approval at University of California, San Francisco, we identified hospitalization encounters for adult (age >=18) COVID-19-positive patients between February 2020 and April 2021. We used a third-party hospital-licensed API to identify the 9-digit zip-code for each patient encounter. We mapped these zip codes to the California ADI database (latest version 2018) from the Neighborhood Atlas2,3, resulting in ADI state-level decile scores of 1 (least disadvantaged) to 10 (most disadvantaged). We excluded patients with zip codes that were out-of-state and/or did not have an associated ADI from the 2018 census-derived ADI scores (i.e. new neighborhoods established after 2018).

To test the significance of the relationship between ADI and LOS, readmission, death, we fit a regression model for each outcome using ADI and covariates listed in Table 1 as regressors. For LOS, we used the log LOS as the outcome variable based on its empirical distribution. For binary outcomes readmission and death, we used a logit link in a binomial regression. We used a semiparametric GAM to model the effects of age and of ADI in each case. Data analysis was performed using Python (3.8.1) and statistical analysis was performed using R (R package mgcv).

Results

Based on our exclusion criteria, we analyzed data from 822 COVID-19-positive admissions (79% of all hospitalized COVID-19 patients). The average age was 60.2 years (SD = 19.5), and 51.6% of patients were male. The average LOS was 12.6 days (SEM = 0.61). The all-cause readmission rate was 13.6% and in-hospital death was 8.8%. Figure 1 shows the distribution of patients within each ADI decile, with most patients in ADI 1-3 (60.1%). Figure 2A shows the average LOS per ADI. Figure 2B shows all-cause readmission rates among the study population’s ADI scores. Notably, ADI of 10 had a 28% readmission rate, which was greater than 2.5-times that of ADI of 1. We found evidence of a relationship between ADI and in-hospital-death (p < 0.001), readmission (p < 0.001), and LOS (p < 0.001).
Table 1. Covariates used in Regression Model

<table>
<thead>
<tr>
<th>ADI*</th>
<th>BMI</th>
<th>Presentation source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>CCI (Charlson Comorbidity Index)</td>
<td>Hospital arrival means</td>
</tr>
<tr>
<td>Sex</td>
<td>Insurance type</td>
<td>Admission year</td>
</tr>
<tr>
<td>Race</td>
<td>Marital status</td>
<td>Primary language</td>
</tr>
<tr>
<td>Ethnicity</td>
<td>Smoking status</td>
<td></td>
</tr>
</tbody>
</table>

*Main outcome of interest

Figure 1. Inpatient Encounter Distribution of ADI

Figure 2. Inpatient Encounters (A) LOS per ADI (B) Readmission Rate per ADI

Conclusion

Our findings suggest a relationship between socioeconomic background and in-hospital outcomes related to COVID-19. Specifically, we find evidence of a relationship between ADI and 30-day readmission, in-hospital death, and hospital LOS in one metropolitan tertiary-care center. Further studies should investigate other confounding factors including specific comorbidities and severity of disease on admission. In the future, Area of Deprivation Index could be incorporated into the electronic health record to help risk-stratify a patient at time of admission, and hospital systems can develop measures to address COVID health inequity related to hospital readmission and death.

References

CovRNN: predicting outcomes of COVID-19 patients on admission using their electronic health records with minimal data processing

Laila Rasmy, MSc1, Masayuki Nigo, M.D2, Bijun Sai Kannadath, M.B.B.S, M.S3, Ziqian Xie, Ph.D1, Bingyu Mao, MA1, Khush Patel, MD1, Yujia Zhou, MSc1, Wanheng Zhang, M.S4, Angela Ross, PhD1, Hua Xu, PhD1, Degui Zhi, PhD1

1School of Biomedical Informatics, UTHealth, Houston, TX; 2McGovern Medical School, UTHealth, Houston, TX; 3College of Medicine, University of Arizona – Phoenix, AZ; 4School of Public Health, UTHealth, Houston, TX.

Introduction

Coronavirus disease (COVID-19) is an infectious disease that emerged in December 2019 and it leads to the death of more than four million patients worldwide by the mid of July 2021. During the peaks of the pandemic waves, many places have reported near-capacity hospital and intensive care unit (ICU) utilization. Accurate prediction of the future clinical trajectories of COVID-19 patients at the time of admission is crucial for clinical decision-making and enables efficient allocation of resources. A survey of the literature till the end of April 2021 showed only two studies1,2 trained their proposed models on more than 20,000 patients. Both of which are based on specific features and needed laborious data preprocessing and feature engineering processes that limit the transferability, reliability, and sustainability of such models.

We developed CovRNN, a collection of deep learning based predictive models for the clinical outcomes of COVID-19 patients including: in-hospital mortality (iMort), need for mechanical ventilation during the stay (mVent), and hospital stay longer than one week (pLOS). CovRNN models were trained on more than 200,000 hospitalized COVID-19 patients to accurately predict patients’ outcomes on admission using readily available structured electronic health records data (EHR) without the need for specific feature selection, feature engineering, or missing values imputation.

Methods

CovRNN utilized a gated recurrent neural networks (RNN) architecture namely gated recurrent unit (GRU) that encodes the temporal nature of patient history including the most recent COVID-19 visit as well as distant events that happened years ago. For iMort and mVent prediction tasks, CovRNN was trained to predict both time-to-event risk scores (survival prediction) as well as the all-time risk scores (binary prediction), to fit different clinical needs for healthcare workers confronting COVID-19.

We kept our data processing to the minimum to facilitate the transferability of our trained models among different datasets. We extracted all patient information on or before the date of their first hospital admission with COVID-19, including demographics, diagnosis, medication, procedures, laboratory results, and observations. We utilized standard terminologies that are readily available in the majority of EHR systems to facilitate interoperability: ICD 9, ICD10, and SNOMED-CT for diagnosis; LOINC and SNOMED-CT for laboratory tests and observations; Multum drug identifiers and categories for medications; and CPT-4, HCPCS, ICD-9 PCs, and ICD-10 PCs for procedures. In cases where a coding system is not used such as Multum codes for medication, pre-existing mapping tools are available that can be used to convert NDC medication codes to corresponding Multum information. For laboratory results, we used clinical interpretations such as “below normal low”, “normal”, or "above normal high", instead of using the actual numerical value. By doing so all our input data became in a categorical format so we can further convert our input into trainable embedding matrices which will learn the features representation that mediates our model robustness to input errors and generalizability.

CovRNN was trained on a large heterogeneous dataset of 243,785 de-identified patients data derived from 85 different health systems and available through the Cerner® Real-World Dataset (CRWD) hosted on the Cerner HealthDataLab. We reported our prediction accuracy and model calibration results on a large heterogeneous held-out test set of 48,781 patients. For external validation, we evaluated the CovRNN on two held-out hospitals' data and compared the model performance against traditional machine learning models logistic regression (LR) and light gradient boosting machine (LGBM) for baseline comparison. We reported the results of subgroup analysis to understand CovRNN performance in different populations and utilized the integrated gradient technique to explain the model predictions.
Results

CovRNN achieved AUC of around 93% for both in-hospital mortality and need for mechanical ventilation binary predictions (vs. around 91.5% and 90% for LGBM and LR respectively) and 86.5% for prediction of LOS > 7 days (vs. 81.7% and 80% for LGBM and LR respectively. For survival prediction, CovRNN achieved a C-index of 86% for mortality and 92.6% for mechanical ventilation. External validation, showed a consistently high prediction accuracy for CovRNN models, as shown in Table 1 and Table 2. Subgroup analysis showed consistent results among different populations based on race, geographical region, and comorbidities as appear in figure 2, while a slight decrease in the model prediction accuracy was observed for older age (65 years or older).

Table 1. Binary classifications models Performance on different CRWD test sets.

<table>
<thead>
<tr>
<th></th>
<th>In-hospital Mortality (iMort)</th>
<th>Mechanical Ventilation (mVent)</th>
<th>Stay &gt; 7 days (pLOS)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LR</td>
<td>LGBM</td>
<td>CovRNN</td>
</tr>
<tr>
<td>Multi- hospital</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test Set</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>48,781</td>
<td>90.3</td>
<td>91.5</td>
<td>93.0</td>
</tr>
<tr>
<td>Hospital 1</td>
<td>3,469</td>
<td>88.8</td>
<td>91.0</td>
</tr>
<tr>
<td>Hospital 2</td>
<td>706</td>
<td>94.6</td>
<td>95.1</td>
</tr>
</tbody>
</table>

Discussion and Conclusion

To the best of our knowledge, CovRNN is the first RNN based model that can accurately predict different COVID-19 patient’s outcomes on admission, using readily available structured EHR in its raw categorical format without any need for specific feature selection, or missing values imputation.

Our trained COVID-19 outcomes prediction models achieve high prediction accuracy comparable to the state-of-the-art in the literature as well as good calibration and lower risk of bias as it was trained and evaluated on large heterogeneous datasets collected from different health systems. Moving forward, our models can be fine-tuned on new data for continuous improvement as recommended by the FDA's good machine learning practice. Furthermore, our framework includes a utility for model predictions explanation to facilitate clinical judgment on the model predictions. For clinicians fighting COVID-19 in the frontline, there are two potential actionable contributions of our work: (i) They can fine-tune our pre-trained model on their local data regardless of the size; (ii) Use our comprehensive model development framework to train a predictive model using their own data. In our work we considered the major factors to establish the feasibility of our model, however, further clinical validation is required.

Acknowledgments L.R. is supported by the UTHealth Innovation for Cancer Prevention Research Training Program Pre-Doctoral Fellowship (CPRIT Grant No. RP160015)

References

Differential Presentation and Delays in Treatment for Acute Myocardial Infarction Associated with Sex and Race/Ethnicity

Harry Reyes Nieva, MAS; ¹ Tony Y. Sun, MA; ¹ Sharon R. Gorman, MA; ¹ Grace Mao; ¹
Noémie Elhadad, PhD
¹ Columbia University, New York, NY;

Introduction

Prior studies concerning disparities in medical intervention following acute myocardial infarction (AMI) have focused mainly on in-hospital mortality and long-term outcomes.¹ ² Given that prior research has shown the importance of early intervention at time of AMI, ideally within 90 minutes of arrival to the emergency department (ED), assessing differences in care practices during ED visits has the potential to produce actionable insights for mitigating disparities in care. Large data networks, such as the Observational Health Data Sciences and Informatics (OHDSI) collaborative, have the potential to support investigation of these epidemiological questions.³ In particular, the electronic health record (EHR), with its highly granular temporal information regarding clinician orders, can provide meaningful data for analysis. Moreover, informatics-based approaches, such as natural language processing of the clinical narrative, can help elucidate differences in disease presentation across patient groups. The aim of this study was to leverage the OHDSI data network to examine differences in presentation of AMI across sex and race/ethnicity, and identify potential disparities in timeliness of treatment.

Methods

We performed a retrospective observational cohort study of ED visits at the Columbia University Irving Medical Center in New York between January 2010 and June 2019 for patients presenting with AMI. We standardized data extracted from the EHR using the OHDSI Observational Medical Outcomes Partnership (OMOP) Common Data Model and identified patients with AMI-related ED visits using a combination of billing codes, EKG findings, laboratory measurements, and coded clinical observations. We excluded any patients with prior history of myocardial infarction. Detailed specifications are available online as an OHDSI ATLAS Cohort Definition.⁴ For each patient, we extracted administrative and EHR data, including ED visit notes. This study was approved by the Institutional Review Board at Columbia University.

Using named entity recognition,⁴ we extracted sign and symptom mentions from clinical narratives. We restricted our search to SNOMED clinical terms associated with the Unified Medical Language System (UMLS) disorder semantic group and the Clinical Observations Recordings and Encoding (CORE) Problem List Subset. We assessed differences in presentation of signs and symptoms among patients using Chi-squared tests with false discovery rate (FDR) correction for multiple comparisons. Using EHR timestamp data, we calculated length of stay and length of time from visit start to orders for first measurement, first medication, and first procedure. Based on these intervals, we fit multivariable Cox proportional-hazards models for in-hospital survival and time to first measurement, medication, and procedure.

Results

We identified 2,286 patients who presented to the ED during the review period with their first diagnosis of AMI. Patients were 54% male and 46% female with a mean age of 66 years (standard deviation [SD], 16) and 72 years (SD, 16), respectively. Among those with available data on race/ethnicity, 30% were Hispanic/Latinx, 16% were Caucasian/White, 11% were African American/Black, and 1% were Asian or Pacific Islander. A larger proportion of women had hypertension (91% vs 88%, p=0.02), were overweight or obese (24% of women vs 17% of men, p<0.001), and had depression (35% vs 31%, p=0.04), while a greater percentage of men had a history of alcohol or substance abuse (3% of women vs 8% of men, p<0.001 and 6% vs 9%, p=0.004, respectively).

Differences in in-hospital survival were not statistically significant for sex (p=0.84) or race/ethnicity (p=0.18). Analysis of clinical documentation reveals that, compared to male patients, female patients were less likely to exhibit diaphoresis and more likely to present with generalized pain, head pain, dyspepsia, epigastric pain, leg pain, esophageal reflux, musculoskeletal pain, malaise/fatigue, neck pain, and impaired cognition (Table 1).

Table 1: Frequencies, percentages, and estimated odds of presenting signs or symptoms by sex*

<table>
<thead>
<tr>
<th>Symptom</th>
<th>Female (N=1043)</th>
<th>Male (N=1243)</th>
<th>Odds Ratio (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>generalized pain</td>
<td>22 (2)</td>
<td>10 (1)</td>
<td>2.83 (1.29-6.21)</td>
</tr>
<tr>
<td>head pain</td>
<td>107 (10)</td>
<td>61 (5)</td>
<td>2.45 (1.73-3.46)</td>
</tr>
<tr>
<td>dyspepsia</td>
<td>47 (5)</td>
<td>34 (3)</td>
<td>1.94 (1.21-3.11)</td>
</tr>
<tr>
<td>epigastric pain</td>
<td>174 (17)</td>
<td>136 (11)</td>
<td>1.72 (1.34-2.20)</td>
</tr>
<tr>
<td>leg pain</td>
<td>218 (21)</td>
<td>189 (15)</td>
<td>1.55 (1.24-1.94)</td>
</tr>
<tr>
<td>esophageal reflux</td>
<td>143 (14)</td>
<td>124 (10)</td>
<td>1.46 (1.13-1.90)</td>
</tr>
<tr>
<td>musculoskeletal pain</td>
<td>439 (42)</td>
<td>439 (35)</td>
<td>1.36 (1.14-1.62)</td>
</tr>
<tr>
<td>malaise, fatigue</td>
<td>475 (46)</td>
<td>479 (39)</td>
<td>1.32 (1.11-1.56)</td>
</tr>
<tr>
<td>neck pain</td>
<td>236 (23)</td>
<td>229 (18)</td>
<td>1.32 (1.07-1.63)</td>
</tr>
<tr>
<td>impaired cognition</td>
<td>573 (55)</td>
<td>617 (50)</td>
<td>1.20 (1.01-1.42)</td>
</tr>
<tr>
<td>diaphoresis</td>
<td>812 (78)</td>
<td>1013 (82)</td>
<td>0.79 (0.65-0.98)</td>
</tr>
</tbody>
</table>

*Ordered from largest to smallest odds ratio.

The analysis of time to different orders in the EHR yielded the following insights. Mean time to first measurement was 0.3 hours (SD, 1.0), first medication was 2.4 hours (SD, 3.1), and first procedure was 11.26 hours (SD, 5.9). We note that the average wait time until first medication and procedure are well beyond the golden response time since onset of symptoms (as they occur more than 2 hours after admission to the ED). Significant differences were found among women who presented with symptoms generally considered atypical for AMI. For example, women presenting with leg pain had longer waits for their first medication compared to men (adjusted Hazard Ratio [aHR]: 0.80; 95% CI: 0.71-0.92; p=0.04), even after correcting for FDR and adjusting for patient age, race/ethnicity, and comorbid conditions. African American and Latinx patients with AMI had longer wait times, on average, than other patients for their first medication (mean difference of 0.39 hours; p=0.04) and first procedure (mean difference of 0.89 hours; p=0.02). They also experienced longer lengths of stay with higher variability (195 hours [SD, 238] vs 194 hours [SD, 191]; p=0.03). Cox proportional-hazards models adjusted for age, sex, and comorbidities indicate that African American and Latinx patients experienced longer time-to-first-measurement (aHR: 0.88; 95% CI: 0.78-0.99; p=0.04) and time-to-first-procedure (aHR: 0.83; 95% CI: 0.73-0.94; p=0.003) compared to all other patients.

Conclusion
The granular data contained in ED documentation (temporality, detailed clinical narratives) enables investigation of differences in outcomes across different patient groups. Our analysis supports consideration of sex-based differences when diagnosing AMI and suggest disparities in time-to-treatment based on sex, racial/ethnic origin, and symptom presentation. The natural language processing analysis of clinical notes confirms that female patients are more likely to experience gastrointestinal symptoms, malaise/fatigue, impaired cognition and types of pain not typically associated with AMI compared to male patients and our analysis suggest that atypical presentation may lead to delays in care. The time-to-event analysis indicates that African American and Latinx patients with AMI have longer wait times and lengths of stay compared to other patients. By leveraging the OHDSI common data model, our analysis is reproducible. Our next steps are to carry out a network study across OHDSI sites with access to EHR documentation. Further work will also build on our analysis to identify patients with atypical presentation who were originally misdiagnosed. Extrapolating from our findings, we hypothesize that atypical presentations would lead to greater misdiagnosis and care delays, particularly among women.

References
Challenges to Public Health Reporting Experienced by Hospitals

Chelsea Richwine, PhD1, Carmelita Marshall, MS, PMP1, Christian Johnson, MPH1, Vaishali Patel PhD MPH1

1Office of the National Coordinator for Health Information Technology, Washington, D.C.

Introduction

Amidst a global pandemic, the need for efficient exchange of public health information between hospitals and public health agencies (PHAs) has never been stronger. To strengthen PHAs’ ability to monitor and respond to disease outbreaks, the Centers for Medicare and Medicaid Services (CMS) require hospitals to report health information to local, state, and federal PHAs as a condition for meeting Promoting Interoperability (PI) program requirements. Understanding challenges hospitals faced related to public health reporting just prior to the pandemic provides insights into their readiness to support PHAs’ abilities to monitor and address the pandemic. This abstract uses nationally representative survey data from the 2019 American Hospital Association (AHA) Information Technology (IT) supplement to describe the number and types of challenges hospitals experienced when submitting health information to public health agencies and how these challenges varied by state and hospital characteristics.

Data Source and Methods

Data come from the 2019 AHA IT Supplement Survey. The survey was administered via mail or a secure online website and was in the field from early January 2020 through the end of June 2020. Target respondents were hospital chief information officers reporting on hospitals’ health IT capabilities and experiences. In this study we analyzed data from a sub-population of 2,645 non-federal acute care hospitals, whose survey response rate was 59%. The sample was further restricted to hospitals that answered the question about challenges experienced when submitting health information to PHAs, resulting in a sample of 2,606 hospitals. Responses were weighted to account for non-response. To understand the extent of public health reporting challenges experienced by hospitals, we assessed the number and types of challenges reported. We also examined whether the types of challenges experienced varied significantly by hospital characteristics (i.e., hospital size, rurality, critical access status, system affiliation) or geographic location.

Results

Number of Challenges Experienced by Hospitals

A majority of hospitals reported experiencing at least one challenge reporting health information to PHAs. Among those who experienced one or more challenges, more than half experienced one or two challenges and less than a quarter experienced three or more challenges (Figure 1).

Figure 1. Number Challenges Experienced by Hospitals, 2019

Source: 2019 AHA IT Supplement Survey

1 Challenge, 28%
2 Challenges, 23%
3 Challenges, 16%
4 or more challenges, 3%

Types of Challenges Experienced by Hospitals

Interface-related issues and a lack of capacity to electronically exchange information were the most common challenges experienced by hospitals (Table 1). About one in five hospitals also reported difficulty extracting relevant information from electronic health records (EHR) or issues exchanging information due to differing vocabulary standards. Approximately 6 percent of hospitals experienced other challenges not listed in Table 1 below.
No HIE participation 4% 51% 15% 38% 17%
HIE participation (23%) 8%* 53% 27%* 41% 16%

Table 1. Types of Challenges Experienced by Hospitals, 2019

| Issues with capacity (e.g., technical, staffing) to electronically exchange information | 50% |
| Interface-related issues (e.g., costs, complexity) make it difficult to send the information | 40% |
| We use different vocabulary standards than the PHA, making it difficult to exchange | 19% |
| Difficulty extracting relevant information from EHR | 17% |
| We don’t know which PHAs our hospital should send info to meet CMS reporting requirements | 4% |

Source: 2019 AHA IT Supplement Survey. ‘Refers to hospitals’ lack of capacity to electronically send information or PHAs’ lack of capacity to electronically receive the information.

Variation in Challenges Experienced by Hospitals

Examine the different types of challenges reported at the state-level reveals substantial variation in public health reporting challenges experienced by hospitals. For instance, half of all hospitals nationally reported issues with the capacity to electronically exchange information. However, the prevalence of the challenge ranges from as low as 17 percent of hospitals in Michigan to as high as 100 percent of hospitals in Rhode Island (Figure 2).

Figure 2. State Variation in Challenges Related to Hospitals’ Lack of Capacity to Exchange Information

Table 2. Variation in Challenges by Hospital Characteristics, 2019

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Don’t know where to send information</th>
<th>Lack capacity</th>
<th>Difficulty extracting information</th>
<th>Interface-related issues</th>
<th>Different vocabulary standards</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small (51%)</td>
<td>6%*</td>
<td>49%</td>
<td>20%*</td>
<td>42%</td>
<td>15%*</td>
</tr>
<tr>
<td>Medium-Large (49%)</td>
<td>2%</td>
<td>51%</td>
<td>13%</td>
<td>38%</td>
<td>23%</td>
</tr>
<tr>
<td>CAH (29%)</td>
<td>8%*</td>
<td>50%</td>
<td>23%*</td>
<td>44%*</td>
<td>10%*</td>
</tr>
<tr>
<td>Non-CAH (71%)</td>
<td>3%</td>
<td>50%</td>
<td>15%</td>
<td>39%</td>
<td>22%</td>
</tr>
<tr>
<td>Rural (40%)</td>
<td>6%*</td>
<td>48%</td>
<td>21%*</td>
<td>44%*</td>
<td>13%*</td>
</tr>
<tr>
<td>Suburban-Urban (60%)</td>
<td>3%</td>
<td>51%</td>
<td>15%</td>
<td>37%</td>
<td>23%</td>
</tr>
<tr>
<td>Independent (34%)</td>
<td>7%*</td>
<td>48%</td>
<td>25%*</td>
<td>44%*</td>
<td>13%*</td>
</tr>
<tr>
<td>System Affiliation (66%)</td>
<td>3%</td>
<td>51%</td>
<td>13%</td>
<td>38%</td>
<td>22%</td>
</tr>
<tr>
<td>HIE participation (23%)</td>
<td>8%*</td>
<td>53%</td>
<td>27%*</td>
<td>41%</td>
<td>16%</td>
</tr>
<tr>
<td>No HIE participation</td>
<td>4%</td>
<td>51%</td>
<td>13%</td>
<td>38%</td>
<td>17%</td>
</tr>
</tbody>
</table>

In addition to geographic variation, the overall number and types of challenges experienced varied by hospital characteristics (Table 2). Small and rural hospitals were more likely to report interface-related issues, difficulty extracting relevant information, and confusion about where to send information. Similar patterns held for independent and critical access hospitals (CAH). Hospitals participating in a health information organization (HIO) were less likely to experience difficulties extracting relevant information compared to non-participating hospitals.

Discussion

Understanding the number and types of challenges experienced by hospitals is critical to developing solutions aimed at improving hospital and PHA capacity to effectively exchange health information. This study shows that just prior to the pandemic a majority of hospitals experienced public health reporting challenges that would impact PHAs’ ability to monitor and address disease outbreaks. These findings are consistent with 2018 survey findings, suggesting these were not emergent issues. Challenges reported by hospitals suggest the need to improve the methods used to exchange information with PHAs. Our findings indicate a potential role for HIOs to support public health by facilitating this exchange. Increased data standardization and bulk reporting could also improve hospitals’ capacity to exchange data with PHAs. Further, variation in challenges by state indicate potential differences in PHA exchange capacity that need to be addressed to support key public health activities related to COVID-19 and future public health emergencies.

References

Signal from the Noise: Quantitative Measures of Conformity and Variability From Process Mining Graphs

Christian C. Rose∗,1, Morteza Noshad∗,2, Jonathan H. Chen2,3
1Department of Emergency Medicine, Stanford University, Stanford, CA
2Stanford Center for Biomedical Informatics Research, Stanford University, Stanford, CA
3Division of Hospital Medicine, Stanford University, Stanford, CA
∗Both authors contributed equally

Introduction

Many life-threatening emergent medical conditions require time-sensitive progression through multiple points of evaluation and management to deliver definitive life-saving interventions. In the case of an acute stroke, blood samples must be collected by a nurse, evaluation must be performed by a physician, the CT scanner must be prepared by a technician and a pharmacist must prepare medications as soon as a stroke has been identified by a radiologist. Studying and improving care processes for these conditions underlies the vision of a learning healthcare system.

Given that these conditions are often complex, requiring multiple providers with differing responsibilities in multiple care settings, it can be difficult to obtain granular detail about how they occur in practice. However, understanding the current process is critical for determining how much variation there is in the process across the organization, where performance problems exist, and where to invest in improvements. Defining current state of processes has hitherto often relied on expert opinion and recall instead of direct observation which can be time-consuming and lack generalizability. Process mining, the method of determining the order of events from an event log, may address these issues.† Process mining can help organizations easily capture workflow information from enterprise systems and provides detailed, data-driven insights about how key processes are being performed. ‡ These logs may illuminate how computer-mediated work, where most of the clinical time may be spent, is really happening - including who did it, how long it takes, and how it deviates from best practice guidelines.

However, a graphical representation alone may not be useful for continuous quality improvement. Hospitals may want to know how much an individual case differed from the standard of care or how variable the process is across their institutions so that they can intervene with educational or decision support initiatives. Here, we utilize process mining of the event log from a common EHR vendor, Epic, at our academic medical center to build a model for tPA management of stroke care. We then utilize the process map to identify the most common pathway (MCP) at our institution. Individual pathways are then compared to the MCP to determine their conformity and the variability of conformity scores across all cases is used to determine variability - comparable to a standard deviation score. It is hoped that the robust mapping and quantitative assessment of critical pathways can then be used to identify and maintain the highest possible quality of care.

Methods

Process mining is based on a set of simple rules to create a graph. In this graph the nodes represent the clinical events, edges represent the subsequent events, edge labels show the relative lag between the events. The graph nodes are time-ordered from beginning to end. For our stroke patients, these nodes began with the emergency department admission and ended with tPA-administration. The edges of the process mining graph connect the pairs of nodes that happen sequentially in the event-log data. They are then given a weight based on the probability that the two events will occur in series over the entire set of possibilities. The most common path is thus the series of edges from beginning to end with the greatest weights.

While the primary application of the proposed process mining graph is to represent the scope of common pathways for a process of interest, it may also be helpful to measure the conformance of the pathways across patients or encounters. Many factors may affect the conformity to a particular pathway. Patients may present for evaluation with different symptoms, from different locations (near or far from the hospital), or the ED may be busy at the time of presentation. A particular encounter can be compared against the rest to better understand how it compares.

We propose the following conformity score using the probability weights of the edges:

\[
C_P = \frac{1}{|E_P|} \sum_{e \in E_P} p_e
\]
where \( P \) represents a path for a patient, \( E_P \) is the set of all of the edges for the patient \( P \), and \( p_e \) are the probability weights for the edge \( e \). A higher conformity score means the patient’s path closer to the expected pathway from the probability of each edge in that pathway, whereas lower conformity scores are associated with encounters that proceeded through a much less probable series of events. The variability score represents the average differences in conformity scores across all cases, which is normalized for all the possible edges.

**Results**

We were able to automatically create a workflow model for the care of acute stroke patients receiving tPA in our hospital by utilizing the event log data captured in our EHR. Our analysis resulted in a graphical representation of the most common pathway (a process map which we could not include here due to space limitations), which progressed according to the series of events anticipated by our hospitals “Code Stroke” guidelines. We were also able to calculate conformity and variability scores for our pathway from our process mapping data, which were a conformity of 0.64 for the MCP and a variance of 0.36 for the entire graph. We identified specific subsets of care which had minimal variability to each other, but poor conformance to the average pathway - which may represent out of hospital code strokes or other practice deviations and settings.

**Discussion**

Simple process measures like “time-to-tPA” have been used to confer the totality of the experience and to set a benchmark for the quality of care despite the complexity of decisions and actions that take place in the setting of emergent conditions. Process mining offers a unique opportunity to unveil the totality of actions taking place from the moment the patient presents by utilizing event log data. Even simple “spaghetti” graphs which result from process mining might help clinicians identify bottlenecks or areas of high variability. They can be used to identify patient or provider clusters (like out of hospital stroke activations) which result in particular outcomes of interest (like faster time to tPA) or to predict the set of future nodes from the current state allowing for streamlined care.

However, such graphs face some limitations. They traditionally weight all pathways equally, limiting their use in identifying the most common pathway or to provide measures for pathway conformity or variability. Such measures could be used to trend the consistency of interventions within an institution and to identify the effect of process or quality improvement initiatives on practice. They can also be used to identify those pathways which result in the highest quality outcome of interest and thus coordinate care accordingly.

This exploratory work represents the first step in utilizing this graphical process to create reproducible measures which may then be used for quality improvement initiatives. Individual cases may be reviewed based on the outcome of the patient to understand if the care provided was close to that which was expected or ”standard” for that institution. Case review committees might review the pathway of those with negative outcomes to understand if they had the expected sequence of events but simply represent an unfortunate, perhaps inevitable outcome. On the other hand, cases with good outcomes but with low conformity scores may represent ”near misses” - an opportunity to intervene for education or training. Furthermore department directors may want to compare variability across sites. High variability may represent an opportunity to intervene to streamline workflows - much like development of code-stroke pathways - which can be compared before and after those interventions.

It is hoped that with a clearer understanding of how care is actually provided, and a reproducible, measurable way to evaluate it we can then begin to describe and thus improve these processes and the patient care associated with them.

**References**


Cyberchondria: Does mHealth Technology Reassure Patients or Promote Hypervigilance? An Observational Cohort Study Following Total Joint Arthroplasty

Benjamin I. Rosner, MD, PhD
1University of California San Francisco, San Francisco, CA

Introduction
Historically, the hypervigilant patient was one who presented for care, sometimes unnecessarily, with heightened concerns about minor signs or symptoms. As the internet democratized the availability of both good and questionable health information, the term “cyberchondria” described a new internet-informed (or misinformation) version of hypervigilance. As mobile health (mHealth) and digital patient engagement (DPE) platforms that asynchronously guide patients through episodes of care over time and offer them regular reminders to watch for red flag symptoms pervade more deeply throughout healthcare, the question is whether such platforms decrease potentially avoidable healthcare encounters as intended, or increase utilization through heightened patient awareness and vigilance.

We describe a two-site study in which a DPE platform guided patients remotely pre and post total joint arthroplasty (TJA) via automated serial push notifications, and examine the impact on 30-day post-discharge emergency department (ED) visit rates. Such platforms have been shown to decrease hospital readmission rates, but we sought to understand whether ED visit rates decrease when patients are guided through longitudinal care plans remotely, or whether reminders about red flag symptoms may lead to unintended excess ED utilization.

Methods
We carried out a retrospective observational cohort study at two clinical sites in Maryland using a DPE platform (HealthLoop) for patients undergoing TJA between April 6, 2017 and October 13, 2018. As the platform was being used for routine clinical care for all patients, the only eligibility requirements were that patients had email or a smartphone and internet or cellular data access. Through the platform, patients received automated pre-operative education, and a 4-week tapering schedule of automated post-operative education, symptom assessments, and reminders of the signs and symptoms of complications including deep vein thrombosis, pulmonary embolism, surgical site infection, hemorrhage, dislocation, and narcotic induced constipation. Patient engagement was measured longitudinally as the ratio of patient interactions with the platform relative to automated, scheduled interactions that patients were prompted to complete. The primary outcome was the rate of 30-day ED visits ascertained by patient self-report. Self-report of this metric through the platform has been validated against claims data and described elsewhere, characterized by an agreement of 0.96 and a kappa statistic of 0.45, and free of non-response bias. We compared these ED visit rates to both a national set of orthopedic practices using the same platform, and to a regional all-Maryland benchmark from the Healthcare Cost and Utilization Project (HCUP) state emergency department database (SEDD) and state inpatient database (SID). Maryland data were selected for this analysis because Maryland is one of a few states for which the SEDD is linked with the SID, enabling patient-specific tracking of healthcare utilization following an index encounter. Fisher’s exact test was used to compare outcomes with \( P<0.05 \) considered significant.

Results
Summary statistics of the available demographics are shown in Table 1. We found that Maryland patients on a DPE platform had 60.7% and 13.4% relative reductions in 30-day ED visit rates after hip (61/149 vs. 884/8662, \( P<0.01 \)), and knee arthroplasty (14/167 vs. 1287/13302, \( P=0.69 \)) respectively, compared to a state HCUP benchmark. (Figure 1) To better understand the drivers of these reductions, we segmented the 30-day ED visits by patients who were treated and admitted and those who were treated and released. For hip and knee arthroplasty, there were greater relative reductions for treat and admit (84.4% relative reduction, \( P<0.02 \); and 78.6% relative reduction, \( P=0.10 \), respectively) than for treat and release (41.1% relative reduction, \( P=0.28 \); and 14.7% relative increase, \( P=0.64 \), respectively), with treat and admit for hip arthroplasty being the only one to reach significance. However, an accentuated effect, with significant reductions in both ED visit types for knee and hip arthroplasty \( (P<0.001) \) was seen in the larger sample sizes associated with the national DPE cohort, although a matching national HCUP
benchmark (with linked SID and SEDD datasets) was not available for a direct geographically matched comparison.

**Conclusion**

The findings from this study suggest that for hip arthroplasty in the two site Maryland cohort, overall ED visits and those ED visits resulting in admissions decreased significantly, without significant change in ED treat and release rates. Furthermore, there were no significant changes for ED visit rates in this cohort following knee arthroplasty. The reason for the differential effect between the two procedure types (i.e. significance for hip metrics but not for knee metrics) is not clear, but could be due to the relatively smaller sample sizes of the Maryland cohort, particularly since consistent trends become more evident with the larger cohort of the national DPE sample. In both cases, hip and knee arthroplasty, ED utilization showed no significant increases. These findings imply that mHealth technologies that increase patient awareness of red flag symptoms, don’t necessarily significantly increase hypervigilance as measured through ED visit rates, and that they may, in some circumstances, decrease these rates as intended. Reduced rates of ED visits may be explained, in part, by better adherence to aftercare instructions resulting in greater complication avoidance.

As mHealth technologies are put into the hands of patients, they have the potential to educate or to exacerbate. Our study demonstrates that it is possible to leverage these systems to elevate patient awareness about red flag symptoms, without increasing cyberchondria that results in significant excess 30-day emergency department utilization.

**References**


**Abstract**

Mobile health technologies are powerful tools for educating patients, but a side effect is their potential to create hypervigilance and increased healthcare utilization. We conducted a retrospective observational study of two clinical sites in Maryland using a digital patient engagement platform, and compared 30-day emergency department visit rates post-arthroplasty against an all Maryland HCUP benchmark. We found no significant increase in 30-day emergency department utilization, and following hip arthroplasty, there was a significant decrease.
The Efficient Inpatient Provider
Wh-EHR to Find Best Practices

Benjamin I. Rosner, MD, PhD¹, Robert Thombley, Wendi Zhao, MD, MHI¹
¹University of California San Francisco, San Francisco, CA

Introduction
Efficient delivery of care is an important prognostic marker of physician well-being and resilience, while the inverse has been associated with burnout and turnover. In recent years, proprietary provider efficiency profiles (PEPs) have been developed by Epic Systems (Verona, WI) that leverage the audit logs (timestamped records of digital actions taken) of the electronic health record (EHR) to describe the relative efficiency of ambulatory care providers, including measures of time spent in the EHR during clinic hours and after hours (“pajama time”). However, little is known about the provider efficiency in the inpatient setting, and the digital factors that distinguish higher efficiency providers from those with lower efficiency. If it is possible to identify high efficiency providers and to understand the behaviors and factors that are associated with efficiency, then best practices could be shared across care settings.

We describe an evaluation of inpatient adult Hospital Medicine providers at a large academic medical center to (1): Identify those with higher and lower levels of average daily work time (a surrogate of efficiency), and (2) Understand intrinsic or extrinsic factors that distinguish between these two groups.

Methods
We performed a 6-month cross-sectional study of adult Hospital Medicine attending physicians on primary inpatient non-teaching services at the University of California San Francisco between January 1, 2019 and June 30, 2019. We used the audit logs of the EHR to characterize physician daily work time, segregating it into the portion carried out locally at the hospital, and the portion carried out remotely (“pajama time”). We then identified the quartiles of physicians with the lowest and highest average daily work times to study factors associated with observed differences.

We developed an ordinary least squares (OLS) regression model with standard errors clustered by provider to examine the degree to which a number of explanatory variables, both intrinsic and extrinsic to the provider explain the differences observed in average daily work time between the highest and lowest provider quartiles. Variables extrinsic to the provider, including daily patient census, daily patient APR-DRG distribution (an encounter level metric of patient illness severity), number of daily discharges to the community and to post-acute care facilities, number of daily minutes of electronic translation services used, and number of daily pages received, were derived from the EHR audit log, translation service logs, and pager logs. Variables intrinsic to the provider included gender and number of years in practice. Student’s two-sided t-test and Fisher’s exact tests were used to compare continuous and categorical variables respectively. P < .05 was considered significant.

Results
Fifty six physicians were captured in our study period, with 58.9% (33/56) female, and an average (SD) of 7.91 (± 2.85) years of practice experience. On this specific non-teaching, attending only service (one of several services on which these

Figure 1. Average daily worktime for all physicians in the study cohort.
hospitalists at our institution practice), they provided care across 1,595 unique patients, totaling 1,073 physician-days. We found that the average daily work time (Figure 1) for physicians ranged from 7.27 to 14.54 hours, representing a 2-fold difference between the extremes ($P < 0.001$). 89.3% of physicians (50/56) logged some remote work time. In comparing the quartile of physicians with lowest and highest average daily work times, we found a significant difference (8.43 vs. 12.24 hours, $P < 0.001$), and no difference in the proportion of physicians between the two quartiles (12/14 in each) who had logged remote work time ($P = 1.0$).

Overall, the OLS model had an adjusted $R^2$ of 0.274, suggesting that the factors considered describe about 27% of the variability in daily work time observed. (Table 1) By way of reference, when studying human behavior, $R^2$ values greater than 0.25 have been characterized by some as large.\(^1\) Residuals were normally distributed and there was no correlation among predictors (maximum absolute value of Pearson correlation coefficient of 0.188, and maximum variance inflation factor of 1.9). The model demonstrated that neither of the intrinsic variables (gender or years in practice) significantly explained variability in worktime between the lowest and highest quartiles. Of the extrinsic variables, neither patient census, APR-DRG, nor the number of community or post-acute care facility discharges done per day reached a level of significance. However, the average number of minutes spent using translation services and the number of pages received were significant ($P = .015$, and $P < .001$ respectively).

### Conclusion

As in the ambulatory setting, efficiency in the inpatient setting may be an important factor associated with physician resilience, clinical outcomes, and quality of care. If we wish to discover and disseminate best practices of highly efficient inpatient providers, we must first identify who they are, and then we must understand the differences between the factors characterizing their practice of care - including total daily work hours - and those of providers who are less efficient.

In our identification of inpatient hospital medicine providers with lower and higher average daily work times, it is noteworthy that many extrinsic factors such as patient census, patient illness severity, and the volume of discharges daily do not significantly explain differences in work time, suggesting that providers are similarly equipped to manage these factors. It is also noteworthy that neither of the intrinsic factors examined, provider gender and years of practice, were significantly associated with differences in work time. Two extrinsic factors, time spent with translation services, and the number of pages received were significant in explaining work time variability. In total, the factors considered explained about 27% of the variability observed in work time, leaving a variety of other potential intrinsic factors to be examined in ongoing research, including, for example, how providers leverage efficiency tools within the EHR (e.g. text shortcuts, order sets, etc.), as well as the daily path navigated through the EHR itself.

### References

Pre- and Intra-COVID-19 Comparison of Nursing Flowsheet Documentation Burden in Acute and Critical Care Units

Sarah Collins Rossetti RN, PhD1, 2, Graham Lowenthal3, Chris Knaplund, MPhil3, Min Jeoung Kang4, 2, RN, PhD, Patricia C. Dykes, RN, PhD3, 4, Sandy Cho, RN, MSN5, Po-Yin Yen, RN, PhD6, Kenrick Cato, RN, PhD2

1Columbia University, Department of Biomedical Informatics, New York, NY; 2Columbia University, School of Nursing, New York, NY; 3Brigham and Women’s Hospital, Boston, MA; 4Harvard Medical School, Boston, MA; 5Newton Wellesley Hospital, Newton, MA, 6Washington University, St. Louis, MO

Introduction
At the outset of the COVID-19 pandemic the need to decrease any burdens on clinicians that took time away from direct patient care or unnecessarily increase risk of exposure was recognized. Decreasing documentation burden was a large focus of these efforts at the local level1 and through a Center for Medicare and Medicaid Services (CMS) waiver.2 These changes to minimum documentation requirements provides a unique natural experiment to study EHR burden. We aimed to study year over year changes in documentation rates and hypothesized that they would decrease in 2020. Our team replicated our methods and compared prior findings quantifying nursing flowsheet documentation from 2017 data3 with additional pre-COVID-19 years (2018 and 2019) and an intra-COVID-19 pandemic year (2020).

Methods
We replicated queries and analytical methods from our prior publication analyzing nursing flowsheet documentation burden of 2017 data.3 We extended our 2017 data set to include two more years of pre-COVID-19 data (2018 and 2019) which allowed for analysis of year over year trends, and included 2020 as the intra-COVID-19 data set for the same 4 acute care units (ACU) (specialty: general medicine) and 2 intensive care units (ICU) (specialty: medical) at a large academic medical center as in our prior study.3 Please see our prior publication for specific methods and rationales that were replicated to extract and calculate rates of documentation per nurse, per patient for 12 hour shifts on our study units, including separating out manually entered data from auto-populated device integrated data entries.3

Prior to the COVID-19 pandemic there was a minimum data set of required data elements, such as vital signs and pain assessment. Our study site activated the hospital’s “disaster documentation” mode (changes in EHR configuration and policy) due to the surge in patient demand from 4/1/2020–7/13/2020. The intent of this “disaster documentation” mode was to reduce the extent and frequency of the data elements required for documentation based on CMS COVID-19 waivers. Our site also provided newly configured EHR “navigator” functionality to streamline the new documentation workflows.1 Two changes had implications for flowsheet documentation: 1) documentation by exception, and 2) reassessment required only once per shift in the ICU and required only as appropriate in the ACU.4 Notably, these are minimum requirements. Nurses may (and should) document as appropriate per patient condition, but that is up to the nurse’s judgement rather than dictated by a universal policy for all patients. Another important study-wide impact during our analysis period was the deferral of all elective procedures and ambulatory appointments beginning 3/16/2020 with a phased recovery period through 7/6/2020 in order to preserve inpatient resources for COVID-19 patients. In order to account for these fluctuating rates of patient admissions, we controlled for census in our documentation rates by calculating per nurse per patient. We calculated a time-series 14 day moving average of nurse flowsheet documentation rates per patient per shift in order to compare year over year changes. We also calculated T-Tests to identify any statistically significant changes in documentation rates year over year. Finally, we calculated data entry sessions and plotted sessions in a histogram to understand when the largest (by percent) documentation period occurred in a shift (figure excluded due to space constraints). Our team is currently performing analyses to also control for patient acuity and these data will be presented at the Annual Symposium. Study IRB approval was obtained.

Results
Our time series analyses indicated overall stable trends of manual flowsheet documentation rates per nurse per patient year over year from 2017 to 2019 (Figure 1), and we found significant changes in flowsheet documentation rates during the COVID-19 pandemic in 2020 (e.g., Apr 2019 vs Apr 2020: t(58)=-2.9735, P=0.0043). Most notably, there were steep increases in documentation rates in March-April 2020 across all of our study units (Figure 1). We also looked for emerging patterns in any specific flowsheet measures that contributed to the increase of the documentation rate but did not identify clear clinical categories driving the increase. Device integrated data entries accounted for
2.51%-2.75% of flowsheet data entries across our sample in 2017, 2.14%-2.41% in 2018, 1.85%-2.05% in 2019, 2.31%-2.53% in 2020. Seventy percent of all manual data entries occur in sessions between 8am - 11am and 3pm - 6pm, with approximately 5% more documentation occurring during the 8am-11am session in 2020.

Discussion
We observed an increase in year over year documentation trends for 2020 with a spike in March to April 2020. Our data also indicates that within a month after “disaster documentation” mode (which is intended to decrease documentation burden) was implemented in April 2020 documentation rates were consistent with the year over year rates of prior years, indicating that factors other than EHR configuration and policy were likely driving nurses decisions to document more. There may be several possible explanations for the increase in documentation, and our study design is not intended to identify explanatory data. However, we do point to a similar pattern that our study team has identified and quantified on an individual nurse level—when a nurse is concerned about a patient he/she increases patient surveillance activities (beyond requirements) of specific flowsheet measures.

It is possible that nurses, as they encountered COVID-19 patients, increased their documentation rates due to clinical concern or uncertainty. This would be consistent with early reports of clinical uncertainty in how to treat COVID-19 patients, and that we know in the early months of the pandemic only the sickest patients were hospitalized. Moreover, during the months when “disaster documentation” mode was active after April 2020, documentation rates only decreased to levels similar to prior years, as opposed to lower than prior years (the intent of “disaster documentation” mode). If the increased documentation we observed in this study is due to nurse’s concern about COVID-19 patients that would have important implications for how to address the problem of documentation burden. In other words, the finding that nurses in aggregate increase their documentation (and presumably increase patient surveillance activities) due to a patient state in the absence of requirements would justify efforts for innovative modalities to better support nurses’ data capture, rather than only a focus on reduction of requirements and data elements within the EHR. Self-imposed documentation burden has also been identified as a product of organizational culture. Nurses’ decisions related to documentation require further investigation. In this study we identified no increase in the amount of automated device-integrated data captured from 2017-2020 and data elements that were sometimes captured using automated device data and other times manually entered, indicating immediate room for improvement in the implementation of device integration. Limitations: This study took place at a single academic medical center. We accounted for census changes, but not for case-mix index/acuity which could also impact documentation. Some trends could also be due to EHR configuration changes unrelated to COVID-19.

Conclusion
We found significant changes in manual flowsheet documentation rates during the COVID-19 pandemic in 2020, including steep increases in documentation rates at a time when decreases may have been expected instead. Future work to explore explanations for why documentation rates changed during 2020 should occur to shed light into nursing practice and decision-making amidst uncertain clinical circumstances and patient states, as well as better formulate the specific barriers facing nursing in the effort to reduce documentation burden.

Funding Acknowledgements: This study was funded by the NINR: 1R01NR016941-01, Communicating Narrative Concerns Entered by RNs (CONCERN). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References
MMRF CureCloud(R) Direct-to-Patient Registry: 
A Workflow and Interim Recruitment Update

Cartik R. Kothari, Ph. D1, Michele Likens, MPA1, Jen Yesil, MS1, 
Leon Rozenblit, Ph. D2, Anne Q. Young, MPH1, Daniel Auclair, Ph. D1, 
Hearn J. Cho, MD, Ph. D1, and Steven E. Labkoff, MD, FACMI, FACP, FAMIA1

1The Multiple Myeloma Research Foundation, Norwalk, CT 
2Prometheus Research, an IQVIA Business, New Haven, CT

Abstract
The recently launched MMRF CureCloud(R) Direct-to-Patient Registry and Research Study aims to generate longitudinal genomic and clinical data from over 5000 myeloma patients and share this data with the patients and their doctors, empowering them to make informed decisions for the management of the disease. More than 950 patients have screened into the CureCloud study in a year since the launch and biosamples have been collected from over 300 patients

Introduction
In July 2020, the Multiple Myeloma Research Foundation (MMRF) launched the CureCloud(R): Direct-to-Patient Registry (CC-DTP) and Research Study (ClinicalTrials.gov Registration Number: NCT03657251) with the objectives of: a) aggregating Electronic Health Records (EHR), genomic data, patient-reported outcomes, immune profiles, and patient surveys from over 5000 myeloma patients in the next five years and b) making these integrated datasets available to patients and their clinicians, enabling them to use this information to make informed decisions for disease management, including targeted therapy selection and clinical trial enrollment.

Methods
STEP 1: The patient registers at the CC-DTP website and answers screening inclusion/exclusion questions. If entry criteria are met, the patient digitally signs the consent to participate in the study, a release of data request, Health Insurance Portability and Accountability Act (HIPAA) Release of Information (ROI) forms and, in the case of a California resident, a California Patient Bill of Rights. The patient also answers a demographic and medical history survey in addition to providing the names of all medical institutions and clinicians where they have received treatment for myeloma. Lastly, the patient is given a blood draw order form to be signed by their primary hematologist/oncologist. The patient is fully enrolled in the study once the form is signed and returned to the MMRF

STEP 2A: Once fully enrolled, a copy of the digitally signed consent and the HIPAA ROI form are transferred to Cota Healthcare, the Electronic Health Records (EHR) data curation partner to begin the process of making EHR requests at each institution where the patient has received treatment. The records, once obtained, are processed for information abstraction.

STEP 2B: Concurrently with Step 2A, the Broad Institute’s CLIA-certified Assay Team uses the CC-DTP shipping portal to ship a blood draw kit to the patient’s home, to await the phlebotomist’s arrival. STE

STEP 2C: Concurrently with Step 2B, an order is sent via an API (Application Programming Interface) to the phlebotomy partner for the MMRF to set up an in-home blood draw. Patients receive messages through the CC-DTP portal alerting them to expect the kit to arrive at their homes and a call from the phlebotomy partner to schedule the heme biopsy needed for genomic sequencing and biobanking. The phlebotomist arrives at the patient’s residence (with Covid-19 precautions in place) at an arranged date to draw four tubes of blood. One tube is sent to the Broad Institute for the 70-gene myeloma-specific panel; the others are sent to the MMRF CureCloud Biobank, where they are kept in liquid nitrogen for future bioassays including immune profiling.

STEP 3: When the sequencing and the variant calls are completed at Broad, the resulting files are transferred, via secure, encrypted channels to the Dana Farber Cancer Institute Department of Pathology where the somatic variant calls are made and confirmed by staff hematopathologists who sign out the genomics report. The genomic report is
returned to the CureCloud workflow engine in PDF and JSON formats. The PDF file is sent to the patient’s primary hematologist.

STEP 4: After the pathology report is returned, the variant calls are read from the JSON files and transmitted to the genomic counseling partner, who then takes those calls and creates a patient-readable report at the 8th-grade reading level. Secure links to the patient report are posted to the CC-DTP Patient Portal with information about clinical trials the patient may be eligible for. Copies of the reports are stored in MMRF Google Buckets as PDF files and in JSON format in the CureCloud database.

![Figure 1: The workflow of the MMRF CureCloud Direct-to-Patient (CC-DTP) Registry](image)

**Results**
In the year since going live, more than 950 patients have screened into the study, of which 727 patients have completed all the steps for enrollment. Blood samples have been collected from 302 patients, from which genomic reports have been generated and shared with 292 patients and their doctors along with their clinical information; the other 2 were set aside for research purposes only (Table 1)

| Table 1: Counts of multiple myeloma patients in MMRF CureCloud (On: August 6, 2021) |
|---------------------------------|--------|
| Signed up on CureCloud homepage and screened in | 953    |
| Completed paperwork and Consent | 727    |
| Doctor approved Blood Draw      | 582    |
| Blood Biopsies Obtained         | 302    |
| Reports generated and shared    | 292    |

**Discussion**
The patient enrollment numbers for CureCloud project are comparable with the numbers for other contemporary observational studies despite the impact of delays due to Covid-19, coordination issues among multiple partners, biospecimen collection issues, and healthcare centers refusing to release patient records for extramural research.

**Conclusions**
The CureCloud study is well on its way towards collecting and integrating heterogeneous data from 5000 patients. The translational analysis workflows developed against these datasets will drive novel research initiatives, which will significantly improve the quality of life and life expectancies of myeloma patients.
Systematic replication of smoking disease associations in the All of Us Research Program

David J. Schlueter, Ph.D.1, Lina Suliman, Ph.D.2, Jacob M. Keaton, Ph.D.1, Tracey M. Ferrara, PhD1, Kyle Webb, M.S.1, Ariel Williams, Ph.D.1, Francis Ratsimbazafy, Ph.D.3, Jun Qian, PhD.2, Lisa Bastarache, M.S.2, Andrea Ramirez, M.D., M.S.4, Joshua C. Denny, M.D., M.S.1

1National Institutes of Health, Bethesda, MD; 2Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN; 3Vanderbilt Institute for Clinical and Translational Research, Vanderbilt University Medical Center, Nashville, TN; 4Department of Medicine, Vanderbilt University Medical Center, Nashville, TN

Introduction
The All of Us Research Program (All of Us)1 is a multi-site, national study with the goal of recruiting at least one million participants to further precision medicine. In this study, we assess the extent to which the currently-available dataset of the All of Us can replicate known findings reported in meta-analyses with a commonly studied environmental exposure, cigarette smoking, as defined in Electronic Health Record (EHR) data and via survey responses completed upon enrollment.

Methods
To assess the association between smoking and phenotypes represented in the All of Us, we implemented three Phenome-Wide Association Studies (PheWAS)2. For the independent variables of interest, we used EHR-defined smoking as well as more granular representations of smoking (ever and current smoking) using participant provided survey results (PPI). We defined ever-smoking exposure from EHR data (‘EHR Ever Smoker’) by identifying all participants with at least two instances on separate calendar days of billing codes related to smoking. Never smokers from EHR data (‘EHR Never Smoker’) were participants with zero occurrences of the billing codes used to define EHR Ever Smokers and at least one other billing code present. For survey-based exposure, we used the All of Us “Lifestyle” survey responses. We used an affirmative response to “Have you smoked at least 100 cigarettes in your entire life” to define “Survey Ever Smoker”, and the branching logic question “Do you now smoke cigarettes every day, some days, or not at all?” with the response “Every day” to designate “Current Smokers.” Participants answering “No” to the 100 cigarettes question were included as “Survey Never Smokers”, and participants skipping the question were excluded. Analyses were adjusted for age at last relevant EHR code, sex at birth, race and ethnicity from survey responses, EHR length as reflected by time between first and last billing code, and number of unique billing codes per record. To benchmark the associations uncovered in the All of Us dataset, we compared phenome-wide significant results to known phenotypic associations assessed in meta-analyses found on PubMed. First, we extracted all smoking and tobacco-related meta-analyses from PubMed (N=1840) using the easyPubMed R package. We excluded studies investigating marijuana, vaping, chewing tobacco etc., which resulted in 538 studies that we matched to known phecodes where possible. We then merged this list to all phenotypes with at least one phenome-wide significant result among the three PheWAS analyses. We ran the analysis using the All of Us Researcher Workbench using Python 3, and will be made available to any approved All of Us researcher.

Results
The PheWAS using EHR smoking, Survey ever, and Survey current smoking returned 657, 489, and 602 phenome-wide significant results, respectively. Figure 1 presents a comparison of all PheWAS effect sizes across the three analyses and crosstabulation of counts for PPI Smoking representation versus EHR smoking among participants who had both. We see that the effect sizes for current smoking are closer to the EHR smoking effect
sizes. For the current-smoking meta-analyses, 63 unique phenotypes were matched to phenome-wide significant results. Of these meta-analytic effects, 53 phenotypes had an effect with a 95% CI excluding zero. Among PheWAS results, 45 phenotypes (84.9%) were concordant (36 of the PPI Current results, 42 of the EHR Smoking results) and 13 phenotypes (24.5%) were discordant (13 of the PPI Current results, 7 of the EHR Smoking results) in the sign of the effect size with at least one meta-analysis effect size (note, several phenotypes had multiple meta-analyses associated with them as can be seen in Figure 2.). For the ever-smoking meta-analyses, 36 unique phenotypes with 32 with CIs excluding zero, were matched to phenome-wide significant results. Of the 32 phenotypes, 28 (87.5%) of the phenotypes in the PheWAS analyses agreed (24 of the PPI Ever results, 25 of the EHR smoking results) with at least one meta-analysis for current smoking. Figure 2 presents forest plots of PheWAS effect sizes compared to their corresponding meta analyses. Here, we can see concordance in the direction of effect size between All of Us and meta-analyses for well-known outcomes such as lung cancer.

Discussion

In this study, we found that the majority of phenome-wide significant phenotypes (that were able to be matched to a meta analysis) found in each of three PheWAS were able to rapidly replicate the direction of at least one previously published meta-analysis effect size. Survey driven smoking provides a more sensitive and detailed measure of smoking behaviour, but both EHR and survey smoking behavior replicated known associations well. Smoking is a well known driver of many human diseases, however, we have shown that such knowledge which has been accumulated over several years of study can be rapidly obtained by a single cohort.

Figure 1. Comparison of PheWAS effect sizes and cross tabulations of smoking representation in All of Us.

Figure 2. Forest plots of PheWAS effect sizes and meta-analysis effect sizes.

References

WA Notify, A COVID-19 Exposure Notifications Tool: Modeling Cases Averted in Washington State
Courtney Segal, BA1, Zhehao Zhang, BS1, Bryant T Karras, MD2, Debra Revere, MLIS, MA1, Gregory Zane, MPH2, Janet G Baseman, PhD, MPH1

1University of Washington, Seattle, WA, USA; 2Washington State Department of Health, Olympia, WA, USA

Abstract
Secure and anonymous smartphone-based exposure notification tools are recently developed public health interventions that aim to reduce COVID-19 transmission and supplement traditional public health infection control efforts, including case investigation and contact tracing. We assessed the impact of Washington State’s exposure notification tool, WA Notify, in mitigating the spread of COVID-19 during its first four months of implementation.

Introduction
In May 2020, the Google|Apple protocol GAEN (Google-Apple Exposure Notifications) was released for use by public health authorities (PHAs) within their jurisdictions. A GAEN-based tool is considered “privacy-preserving” in that user’s phones log their exposure history without requiring geospatial location tracking or allowing access to personal data. GAEN-based tools operate as decentralized systems. The Association of Public Health Laboratories (APHL) hosts both a multi-tenant verification (Google) and a key server (Microsoft) supporting interoperable GAEN implementations. Once added to a smartphone, exposure notification tools utilize Bluetooth technology to determine digital proximity between devices and exchange random, ephemeral cryptographic keys when two smartphones are within the physical distance and duration of time specified by a given PHA. A user who tests positive for COVID-19 can anonymously notify others through the tool by voluntarily entering a verification code which then allows them to upload keys to the APHL server. Other users whose proximity and duration of exposure to the index case match specifications set by the PHA are alerted with an exposure notification message. WA Notify1, using the GAEN EN Express solution launched in Washington (WA) State, has been installed on more than two million devices, representing approximately 33% of the adult population. Exposure notification technology represents a potentially disruptive public health informatics strategy for pandemic control. Recent modelling and simulation evaluations have demonstrated the epidemiological value of exposure notification tools in improving contact identification, particularly of contacts unknown to the index case, and slowing secondary transmission of infection2,3. The preliminary evaluation of WA Notify described here seeks to answer the question: To what extent did WA Notify avert new COVID-19 cases in Washington during the first four months of its use?

Methods
Empirical validation of WA Notify’s epidemiological impact and value as a non-pharmaceutical intervention is limited by the anonymity of users and data privacy. Within these constraints, the approach described herein were designed to leverage aggregate metrics across disparate sources, including de-identified WA State DOH contact tracing data, aggregated APHL verification code server metrics, the MITRE Exposure Notifications Private Analytics Dashboard Data, and a survey among exposure notification recipients. To estimate the number of cases averted by WA Notify, a modeling approach developed to evaluate the NHS COVID-19 app was adapted2. This model is essentially a product of five terms: (i) number of notifications generated, (ii) secondary attack rate, i.e., the probability that notified individuals go on to become cases, (iii) expected fraction of transmissions preventable by strict quarantine of an infectious individual after a notification, (iv) actual adherence to quarantine, and (v) expected size of the full transmission chain that would be originated by the contact if they had not been notified.

Results
The mathematical modeling to estimate number of cases averted used variations of the parameter calculations, presented in Figure 1. The chosen parameters used reflect team and WA State DOH congruence regarding most realistic variation in SAR and quarantine behavior within the bounds of the modeling approach. Figure 1 illustrates the estimated number of COVID-19 secondary transmissions averted at varying levels of secondary attack rate and quarantine effectiveness.
Between 11/30/2020 and 03/31/2021 modeling estimates that:

- Assuming 5.1% SAR and 53% quarantine effectiveness, 2636 cases are averted.
- Assuming 6.659% SAR and 53% quarantine effectiveness, 3439 cases are averted.
- Assuming 12.085% SAR and 53% quarantine effectiveness, 6240 cases are averted.
- Assuming 12.085% SAR and 64% quarantine effectiveness, 7536 cases are averted.
- Assuming 13.706% SAR and 64% quarantine effectiveness, 8547 cases are averted.

After conducting a sensitivity analysis, the estimated number of COVID-19 cases averted is 5,500 (central 95% range of sensitivity analysis 2,800-8,200). Applying an estimated case fatality of 1.4% in WA State to the cases averted estimates, WA Notify saved 40-115 lives between November 30, 2020 and March 31, 2021.

Conclusion

GAEN-based exposure notifications are just one public health tool among many (vaccines, contact tracing, masks, hygiene, social distancing, occupancy policies) that should be leveraged to contain the spread of COVID-19. With the expanding evidence-base for their effectiveness, integrating informatics tools like WA Notify into public health practice alongside other established interventions will be helpful for epidemic control in the future. As new variants emerge and non-essential travel bans are lifted, there is uncertainty about the future of the COVID-19 pandemic and a need to maintain effective pandemic control strategies. As we emerge from the COVID-19 pandemic, the application of digital proximity detection and anonymous exposure notifications may shift to other public health emergencies in the future and strengthen community-clinical engagement with added functionality (e.g., requesting a test or reporting vaccination status) and informatics research.

References

Enhancing Clinical Data Analysis by Explaining Interaction Effects Between Covariates in Deep Neural Network Models

Yijun Shao, PhD1,2, Ali Ahmed, MD, MPH1,2,3, Edward Zamrini, MD4, Yan Cheng, PhD1,2, Qing Zeng-Treitler, PhD1,2

1George Washington University, Washington, DC, USA; 2Washington DC VA Medical Center, Washington, DC, USA; 3Georgetown University, Washington, DC, USA; 4University of Utah, Salt Lake City, UT, USA

Introduction

Deep neural networks (DNNs) have attracted much attention in recent years, and application of DNNs to medicine and healthcare quickly followed and has shown some success.1 Despite their increasing popularity, DNN models are very difficult to explain or understand, which makes them harder to adopt by clinicians and clinical researchers. Current explanation approaches include Attention Mechanism, Local Interpretable Model-Agnostic Explanations (LIME), and Layer-wise Relevance Propagation (LRP).4 However, model-specific approaches are not easily generalizable, and model agnostic approaches like the Shapley value5 can be computationally very expensive. Covariate interactions are commonly observed in healthcare studies. It is natural to ask whether DNN models can capture interaction effects of two variables and how to measure the interaction effects. In this study, we develop a quantitative method called "interaction scores" to measure covariate interactions and apply it to two DNN models trained on two datasets respectively: a simulated dataset and a real-world clinical dataset.

Methods

Consider a DNN with one input layer of n nodes, one output layer with one node, and a number of hidden layers with various number of nodes each. The activation function for the output layer is assumed to be the sigmoid function \( \sigma(x) = 1/(1 + e^{-x}) \) so that the output is a single value between 0 and 1, which is commonly used for predicting binary outcomes such as mortality. Let \( p = F(x_1, ..., x_n) \) denote the final DNN model, where \( x_1, ..., x_n \) are the n predictor variables corresponding to the n nodes of the input layer. Define \( f(x_1, ..., x_n) = \logit(F(x_1, ..., x_n)) \), where \( \logit(x) = \log \frac{x}{1-x} \) is the inverse function of the sigmoid function \( \sigma \). For each variable \( x_i \), we fix a reference value \( x_i^r \).

For two variables \( x_i \) and \( x_j \), we define the individual-level interaction score between them as

\[
\text{Interaction score} = \frac{f(\cdot x_i \cdot x_j \cdot \cdot \cdot) - f(\cdot \cdot \cdot) - f(\cdot x_i \cdot \cdot \cdot) + \cdot f(\cdot \cdot \cdot)}{x_i^r - x_i^r} \text{ for each individual subject with } x_i \neq x_i^r \text{ and } x_j \neq x_j^r.
\]

The population-level interaction score is similarly defined as the mean of all individual-level interaction scores. We apply the interaction scores to two datasets. The first dataset is created from simulation, and the second contains clinical data from the real world.

Simulation Data. The simulation uses 100 variables \( x_1, x_2, ..., x_{100} \) as predictors and a variable \( z \) for binary outcomes 0 and 1 respectively, which satisfies a nonlinear relationship as \( \logit(P(z = 1)) = \beta_0 + \sum_{i=1}^{100} \beta_i x_i + \sum_{m=1}^{20} \gamma_m x_{jm}^2 + \sum_{m=1}^{20} \theta_m x_{jm} x_{km} \), where \( \{i_m\}_{m=1}^{20}, \{j_m\}_{m=1}^{20} \text{ and } \{k_m\}_{m=1}^{20} \) are 3 randomly sampled out of the index set \( \{1, 2, ..., 100\} \) with \( \{j_m\} \cap \{k_m\} = \emptyset \). We generate a set of 50,000 samples, which is the simulation data.

Real-world Data. The dataset is derived from a cohort developed by Cheng et al.6 for the study of Alzheimer’s disease and related dementia (ADRD) and contains 500,000 patients of only two races: black and white. The outcome is the ADRD diagnosis within 10 years of follow-up.

DNN Architectures. The DNN model for the simulation data will be a multiple layer perceptron (MLP) with 10 hidden layers, and the DNN model for the Real-world data will be an MLP with 8 hidden layers.

Results

The DNN model trained on the simulation data achieves a testing AUC of 0.945. The comparison of the interaction scores calculated based on the true model and those based on the true relationship is shown in Table 1. The DNN model trained on the simulation data achieves a testing AUC of 0.945. The comparison of the interaction scores calculated based on the DNN model and those based on the true relationship is shown in Table 1.
Table 1. Comparison of the computed (population-level) interaction scores with the true interaction coefficients in the nonlinear relationship underlying the simulation data.

<table>
<thead>
<tr>
<th>Comparison Metric</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson's correlation</td>
<td>0.86</td>
</tr>
<tr>
<td>Spearman's correlation</td>
<td>0.98</td>
</tr>
<tr>
<td>Sign agreement</td>
<td>1.0</td>
</tr>
</tbody>
</table>

The DNN model trained on the real-world data achieves an AUC of 0.740. We calculate the interaction scores and show the top 5 interactions of the largest-in-magnitude interaction scores in Table 2.

Table 2. List of variable pairs involving race, and their (population-level) interaction scores on the real-world data.

<table>
<thead>
<tr>
<th>Variable 1</th>
<th>Variable 2</th>
<th>Interaction Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGE</td>
<td>RACE: Black (vs. White)</td>
<td>0.6891</td>
</tr>
<tr>
<td>BMI</td>
<td>RACE: Black (vs. White)</td>
<td>-0.5328</td>
</tr>
<tr>
<td>RACE: Black (vs. White)</td>
<td>ETHNICITY: Hispanic (vs. Non-Hispanic)</td>
<td>-0.2355</td>
</tr>
<tr>
<td>ATHRITIS</td>
<td>RACE: Black (vs. White)</td>
<td>-0.1875</td>
</tr>
<tr>
<td>AFIB</td>
<td>RACE: Black (vs. White)</td>
<td>0.1851</td>
</tr>
</tbody>
</table>

Discussion and Conclusion

Significance: DNN model’s ability to capture unknown or undefined interaction present in nonlinear relationships is one of its strengths. Such captured interactions may reveal important underlying relationships (e.g. drug-drug, race-risk factor, gene-environment) in biomedical data. Prior explanation of DNN focused on individual variable’s contribution to outcomes. We proposed a novel method to measure interactions captured by DNN models in this study. Since it is difficult to validate the authenticity of the interactions assessed by our method in clinical data, we first utilized a simulated dataset. The results showed a high level of agreement between the interaction scores and the interaction coefficients underlying the simulation data. We then applied the method to an ADRD dataset, representing real-world data. The interactions between different ADRD risk factors is not well known in general. The DNN model and the interaction score make it possible to study and better understand the interactions of the ADRD risk factors.

Implication: The ability to quantitatively assess interactions allows for an additional means to explain DNN models. This is particularly important because DNN models are often not linear. Assessing individual variable’s importance to or impact on the outcome, of course, is very helpful. Interaction score can complement the explanation provided on individual variables. The assessment of interactions can generate new hypotheses for future exploration. For example, the interaction between race and a number of comorbidities in the context of incident ADRD is not well documented in literature and requires further examination. A better understanding of the race-risk factor interaction could lead to improvement in care of African Americans. It could also support a more personalized medicine approach in ADRD prevention.

Acknowledgment

The authors thank the U.S. Department of Veteran Affairs for providing data and analytical software support. This work was in part supported by the grants NIH 1UL1TR001876-01, VA I01 HX001145-02, NIH 1RF1AG069121-01.

References

5. Shapley LS. Notes on the n-Person Game -- II: The Value of an n-Person Game. 1951. Santa Monica, Calif: RAND Corporation.

1510
Prescribing Pharmacogenomics Testing: Analyzing the Acceptance of Healthcare Providers through a Survey Study
Mohit M. Sharma MPH¹, Yonaka Harris MS¹, Yiye Zhang PhD¹, Jyotishman Pathak PhD¹
¹Weill Cornell Medicine, New York, NY

Introduction
Pharmacogenomics (PGx) is a branch of research that examines the effect of genes on the response of an individual to medications. Considered as a part of precision medicine, it is intended to assist the doctors in choosing the medications which are best suited for patients, thus helping them treat each patient individually (1). Presence of acceptance barriers is causing a significant impediment in the way of harnessing the scientific advancement in this area into clinical practice. To investigate the barriers and opportunities in the adoption of PGx testing for the selection of antidepressant, we conducted a survey of healthcare providers’ current knowledge of and experience with PGx testing and their perception of its utility in routine clinical care (2).

Methods
We distributed electronic cross-sectional surveys using Qualtrics (Qualtrics, Provo, UT) to anonymous primary care providers (N=340) in Weill Cornell Medicine/New-York Presbyterian campus. Participants were offered a gift card of $25 as compensation for their participation in the survey. They were provided with an information sheet which contained details on how the survey answers will be used, the risks/benefits associated with the survey, the confidential nature of the data, and assured that the participation of subjects was voluntary. We adapted the survey present in Haga et al. (3). All the participants provided demographic details at a baseline level. Likert-type questions in assessing their understanding, knowledge, experience, and preferences regarding Pharmacogenomics testing were included in the survey. The response choices include - neither agree nor disagree, somewhat agree, somewhat disagree, strongly agree, and strongly disagree. This study was approved by the WCM internal review board (IRB).

Results
As of August 2020, 158 responses were collected. The response rate is 18% and the completion rate is about 90%. Male healthcare providers constituted 55.4% of the total respondents and females were 44.6%. Likert-type responses were collapsed (for example, strongly disagree and somewhat disagree) to estimate the proportion of providers in agreement or disagreement. In general, 67 out of 134 of the providers agreed to feeling well informed about genetic testing in general while only 37.13% (50 out of 134) expressed their disagreements. Moreover, 60 providers out of 130 (46.15%) did not feel well informed about the role of PGx testing in choosing a psychotropic substance, whereas 55 out of 130 (42.30%) agreed to being well informed.

We observed that younger healthcare providers were less comfortable in ordering a genetic test to diagnose a disease, with 53.2% of them under 30, and 70% of them in the age group 31-40 respectively. Across the specialty, 31 out of 65 (47.69 %) of psychiatry providers were not comfortable in ordering a genetic test, whereas 58.62 % i.e 17 out of 29 Internal Medicine providers answered the same. Seventy six out of 133 healthcare providers (57.14 %) did not feel that their genetic training had sufficiently prepared them to order genetic tests and utilize the results clinically, with 21 of them (15.79%) not being able to decide on a conclusive response.

To evaluate the experience of ordering genetic tests, we asked them how often in the past year had they ordered genetic testing for either disease susceptibility or diagnosis. Seventy out of 133 (52.63%) of the providers never ordered genetic testing for either disease susceptibility or diagnosis whereas 35.54% (47 out of 133) of them had ordered only 1-2 times during the year. Three providers each have ordered for 11-25 times per year and more than 25 times per year respectively. This response reflects about the lack of experience amongst all the providers in ordering genetic tests in general. Across specialty, we observed that the highest number of providers not having any experience in prescribing genetic tests belonged to Psychiatry (31.57%) followed by Internal Medicine (11.27%).

Providers were asked that if a comparable non-genetic test (such as enzyme, plasma level, or protein assay) were also available, would they be more likely to order such a test to predict drug safety or efficacy for an individual patient (as opposed to the genetic test). It was found that although 71 out of 132 (83.53%) were of the opinion to order such non-
genetic tests, 47 out of 132 (35.61%) of the providers were unsure, with only a handful of them i.e 14 (16.47%) prefer not to order such a test.

Majority of the providers (53.04%) believed that PGx testing is or will become a valuable tool to predict risk of adverse events or the likelihood of treatment response for psychotropic medications while 35 responses out of 130 (26.92%) were not sure of their beliefs. Moreover, 56 out of 129 (43.15%) of the providers could not give a conclusive answer when asked if they believe that there is a strong evidence base of PGx testing for psychotropic medications, followed by 42 of them (32.55%) who responded negatively. Only 24% of the responders agreed. Fifty-nine out of 129 (46%) healthcare providers felt that the PGx testing will be beneficial for their patients with Major Depressive Disorder, however there were 42 providers (32.5%) who could not decide about its benefit.

Conclusion

Our findings suggest that despite the healthcare providers having knowledge of the PGx testing and understanding the benefits, there are significant concerns and barriers in the adoption of PGx testing in standard clinical practice. Some of the responses were heterogeneous in nature while in some questions, lack of confidence in expressing clarity of opinion was observed. Our findings are consistent with the ones obtained in previous studies where physicians received sporadic education on the topic of PGx and thus felt underconfident for ordering the tests, even when they expressed almost universal acceptance of the concept of PGx. This underscores the need of having rigorous implementation initiatives and more frequent education programs. The next phase of the analysis includes regression analysis to further investigate the multivariate relationships.

References

2. Hull LE, Lynch KG, Ph D, Oslin DW. VA Primary Care and Mental Health Providers ’ Comfort with Genetic Testing : Survey Results from the PRIME Care Study. 2019;799–801.
Outpatient Portal Use in Pregnant Women for Blood Glucose Management

Priti Singh¹, PhD, MS; Pallavi Jonnalagadda¹, MBBS, DrPH; Evan Morgan², BA; Naleef Fareed¹,², PhD, MBA

¹The Center for the Advancement of Team Science, Analytics, and Systems Thinking in Health Services and Implementation Science Research (CATALYST), ²The Ohio State University, Columbus, OH; Department of Biomedical Informatics, College of Medicine, The Ohio State University, Columbus, OH

Keywords: patient portals, technology use, health information technology, diabetes in pregnancy

Introduction

Health information technologies (HIT) have transformed the health care industry by reducing the patient-provider communication time, increasing accessibility to health information and providing individuals with tools to manage their health. The integration of Electronic Health Record (EHR) with HIT has established new avenues for care delivery and care co-ordination including patient portals. A patient portal is a patient facing health technology tool that provides secure access to personal health information including medication list, laboratory results, upcoming appointments, encounter summaries, billing and more. Government initiatives, and greater accessibility to the internet and internet-based devices have enhanced portals use in recent past. Preliminary evidence suggests positive outcomes of portal use such as greater patient engagement, improved clinical decision making, higher levels of patient satisfaction and improved health outcomes such as medication adherence ¹. Patients with chronic conditions, who need to manage care across multiple sites and providers, especially benefit from portal use due to reasons such as convenience, patients’ own insight and support of care continuity ². Extant studies have examined portal usage in nonpregnant adults, few have focused on pregnant women. Pregnancy is a time for continuous care with higher demands of patient-provider interaction and health monitoring. Given the potential benefits of portal use and frequent contact with the health care system, it is vital to examine patient portal use during pregnancy. Particularly, among pregnant women with existing comorbidities like diabetes requiring frequent clinical encounters and laboratory tests. Diabetes is among the most commonly encountered conditions in pregnancy, affecting up to 10% of pregnant women in the US. Yet, few studies have examined portal usage and their association with blood glucose levels. Analysis of log files generated from pregnant women’s use of the portal present an opportunity to understand technology engagement in this understudied population. Knowledge of patterns of utilization of patient portal and their association with clinical outcomes can inform the development of better health care delivery strategies. To this end, we examined the association between demographic and clinical covariates, blood glucose levels and outpatient portal use in a sample of pregnant women receiving care in the Obstetrics and Gynecology (OB/GYN) department and the Division of Maternal Fetal Medicine (MFM) at our academic medical center (AMC). We hypothesized that women with elevated blood glucose levels will have higher use of the patient portal, as measured by the number of times all the portal functions are used.

Methods

The study followed a retrospective cohort design, wherein we extracted server audit log files from the outpatient portal, MyChart (MC) in use at our AMC from 2016 – 2020. Women eligible for inclusion must have been 18 years or older, has a pregnancy episode and were cared for by our AMC’s OB/GYN and MFM providers. There were 17,224 patients with active MC accounts, who had a prenatal visit (initial or follow-up) and a pregnancy episode during the study period. Patients were excluded from the study if they did not have a prenatal visit with OSUWMC OB/GYN and MFM providers or were not pregnant during the study period. Our AMC MC offers the following functions: Messaging (links to messaging center, letters to the patient, prescription refill option); Visits (list of past and upcoming visits); My Record (list of medications, allergies; medical history, immunizations; test results and health summary; preventive care and a summary of plan of care); Medical Tools (share medical records with other services; participate in research studies; and connect tracking devices); Billing (account summary, payment); Resources (terms and
conditions; patient education; and frequently asked questions); Proxy (request or renew proxy); Preferences (personal and security settings; notification preferences); and Custom (miscellaneous).

Analysis and Results

We used a negative binomial regression to model the association between portal usage and blood glucose levels adjusting for episode of pregnancy (1st pregnancy, 2nd pregnancy, and so on), age at first clinical encounter, Charlson Comorbidity Index, an area-level measure of deprivation (ADI) based on Zip Code, pregnancy risk (high/normal), race, and BMI at the initial encounter. Geometric mean was obtained for glucose values collected throughout pregnancy. The ADI scores were obtained from freely available Neighborhood Atlas website and measures the neighborhood disadvantage with the public, including educational institutions, health systems, not-for-profit organizations, and government agencies. Those with multiple pregnancy, diabetes, high blood pressure, genetic conditions, history of premature birth, preeclampsia, advanced maternal age, or any condition requiring high-risk care or fetal treatment were classified as high pregnancy risk. Our model included an offset for the duration of portal use. Findings from fully-adjusted models are reported as incidence rate ratios (IRR) and 95% confidence intervals (CI) significant at a p-value <0.05. Overall, patients used the Visits feature the most, followed by My Record. No activity was observed for the Resources feature. For every 10 unit increase in blood glucose, portal usage was 3% higher after adjustment (IRR: 1.003, 95% CI:1.002, 1.004). Women living in areas with the high ADI rank (most deprived) (quantile 5) had lower portal usage as compared to those living in least deprived areas (quantile 1) (IRR: 0.80, 95% CI: 0.73, 0.88). Overall, adjusted portal usage was almost 30% lower in Black women as compared to White (IRR 0.73, 95% CI 0.68, 0.78). Patients with high-risk pregnancy had significantly higher use of the Visits, MyRecord, and Billing features, as well as, the overall portal usage than normal risk patients, as indicated in the figure.

Conclusion

We found differences in outpatient portal use based on disease severity among pregnant women. Further, women living in areas of high opportunity had significantly higher use of the outpatient portal. Future work should focus on examining the association between outpatient portal use and clinical outcomes of diabetes management like glycemic control and medication adherence. Our findings have implications for care providers, who can provide encouragement, training, or navigational assistance to pregnant women to use outpatient portals to better manage their health.

References

User Evaluation of Interactive Longitudinal PRO Visualizations Designed by Prostate Cancer Survivors with Limited Graph Literacy

Lauren Snyder MPH1, Daniel Phan RN1, Sarah E. Connor MPH2, Sheba George PhD3, Kristen C. Williams MA2, Jefersson Villatoro2, Ayan Sara4, Lorna Kwan MPH2, Nick Reid MHI1, John Gore MD, MPH3, Mark S. Litwin MD, MPH3, Andrea L. Hartzler PhD1
1School of Medicine, University of Washington, Seattle, WA; 2Department of Urology and 3Department of Community Health Sciences, University of California, Los Angeles, CA

Introduction

Visualizing patient-reported outcomes (PRO) over time has become a common strategy to help patients track their health and engage in healthcare. However, traditional line graphs and bar charts might not be understood by people with limited numeracy and graph literacy. Prostate cancer is the most common solid malignancy in men. Although it causes over 33,000 deaths annually in the United States, many men live for many years with indolent low risk disease.

Prostate cancer survivors with limited graph literacy may also have low socioeconomic status, limited health literacy, and experience other health inequities and resulting poor health outcomes associated with racial/ethnic minority group membership. Although White, college educated prostate cancer survivors prefer PRO visualizations in line graph and bar chart formats, these formats may not be accessible, acceptable or usable to all survivors.

To investigate longitudinal PRO visualizations codesigned with prostate cancer survivors with limited graph literacy, we conducted a series of design studies. We engaged survivors from a state-funded program providing free prostate cancer care to low-income Californians at Martin Luther King Jr. Outpatient Clinic, a safety net community clinic. Our previous work describes our process to understand user needs and to co-design 3 longitudinal PRO visualizations with our Patient Advisory Board of 6 survivors where numerous designs were assessed to identify key features needed for a PRO visualization acceptable to this population. In the present study, we engage prostate cancer survivors who have limited literacy to assess the 3 visualizations (Figure 1) compared with a bar chart that combines emojis (“emoji bar”) as a comparator from the literature. We use virtual methods to engage representative users from this hard to reach, diverse population.

Figure 1. Interactive PRO visualizations that illustrate urinary function scores from the Expanded Prostate Cancer Index Composite over 48 months. The Meter (top left) includes buttons to change the time point; in the Words, (top right) time can be changed on the axis or by arrow; and the Comic (bottom) includes audio of the conversation.
Method

The objectives of our user evaluation were to measure 1) the accessibility, using the metrics of preference, comprehension, and utility of the PRO visualizations, and 2) assess the usability of interactive versions in the subset of participants who are able to use Zoom. To accomplish these objectives, we use mixed-methods to collect data through remote phone/Zoom interviews during the Covid-19 pandemic in February to April 2021. A critical component of this work is engaging prostate cancer survivors who are representative of our target audience, namely prostate cancer survivors with limited graph literacy. We recruited English-speaking survivors with high school education or lower. We characterized literacy with the Newest Vital Sign (NVS) and Short Graph Literacy Scale (GLS). We mailed participants paper versions prior to the interview and afterwards sent them a $50 gift card.

To assess accessibility we measured comprehension (correct/incorrect), utility ratings (1=“not at all helpful” to 4=“extremely helpful”), and rank order preference among the four visualizations (i.e., meter, word, comic, emoji bar). We also collected qualitative feedback to assess acceptability. For participants with access to a computer and Zoom, we assessed the usability of interactive, web-based versions of our three visualizations through task analysis (e.g., participants interacted with the visualization to determine urinary function scores at different time point and the System Usability Scale (SUS). Interviews were audio recorded and transcribed for qualitative analysis.

Results

We conducted 18 interviews with prostate cancer survivors, 4 of whom were willing and able to complete the usability assessment via Zoom. The mean age of participants was 69 (range 61-77), and we purposively sampled for racial/ethnic diversity, including 50% African American and 16.7% Hispanic participants. 89% of participants had a low graph literacy score as measured by the GLS and 78% had low literacy as measured by the NVS. The average comprehension of the Meter was 74%; the Words was 68.5%; the Comic was 61.1%; and the emoji was 79.6%. We used the constructs of helpfulness, satisfaction, and confidence to measure utility. Participants ranked the Meter and Emoji prototypes highest for helpfulness (3.1 and 3.2 out of 4, respectively) and satisfaction (4.3 and 4.5 out of 5, respectively). For confidence, all prototypes received nearly the same score of 3.1 (+/-1) out of 4. Most frequently ranked the Meter and Emoji as their first or second most preferred visualization. In the usability testing with 4 participants, the Meter and Words scored better on the SUS than the Comic; we heard from participants that the Comic made light of a serious health condition and was not relatable or appropriate. This feedback was in stark contrast to the input of our PAB during the design process who highly favored the comic and spent a great deal of time helping us refine and improve the PRO visualization. One participant even audio recorded their voice to provide a narration functionality to the online version. These differences speak to the need to create visualizations that are evaluated by a sample of end users who have not been closely involved in the design.

Conclusion

Findings build on previous research to address the specific needs of prostate cancer survivors with limited graph literacy. We used mixed-methods with standardized instruments that have been adapted for virtual administration. By collaborating directly with our survivors throughout the design process, we increased the likelihood that our visualizations are reflective of and responsive to the needs of our specific population, and ultimately improve accessibility, acceptability and usability of longitudinal PRO visualizations to all survivors.

References

5. Snyder LE, Saraf AA, Perez RC, et al. Visualization co-design with prostate cancer survivors who have limited graph literacy. Visual Analytics in HealthCare Workshop; 2020 Nov 14-18; Chicago, IL.
A Scoping Review of Informatics Research for Clinical Practice Variation

Sunghwan Sohn, PhD1, Sungrim Moon, PhD1, Larry J Prokop, MLS3
Victor M Montori, MD, MS4, Jungwei Fan, PhD1,2*
1Department of Artificial Intelligence & Informatics, 2Kern Center for the Science of Health Care Delivery, 3Mayo Clinic Libraries, 4Department of Internal Medicine
Mayo Clinic, Rochester, Minnesota, USA

Introduction
Clinical practice involves reasoning and decision-making that are affected by interlaced factors from the patient, clinician, and environment – all contributing to potential variation. Clinical practice variation (CPV) stands for a common phenomenon that a specific clinical task is performed differently across providers, regions, or times. For example, inter-provider variation in prescribing prophylactic antibiotics prior to urologic procedures. The issue is becoming ever significant in the face of our aging, multi-morbid population susceptible to disease-disease, drug-drug, and drug-disease interactions, for which credible guidelines are lacking and clinicians frequently need to make autonomous judgment based on their beliefs, instincts, or even incentives. CPV is an established area with decades of scholarly work in health services research, a discipline that has evolved with medical informatics in parallel and shares increasing methodological interests. Without attempting to compare the two disciplines, we believe a dedicated survey of informatics research in CPV will be extremely helpful to facilitate a cross-disciplinary dialogue. Accordingly, we conducted a scoping review of CPV studies published within informatics journals and conferences.

Methods
We targeted on CPV research published in major informatics journals or proceedings. The Ovid integrated database was searched on October 27, 2020, which covered EBM Reviews - Cochrane Central Register of Controlled Trials, EBM Reviews - Cochrane Database of Systematic Reviews, Embase, Ovid MEDLINE and Epub Ahead of Print, In-Process & Other Non-Indexed Citations and Daily. An additional small set of articles were collected by serendipitous browsing or referral from colleagues. Based on the search hits, two annotators (SS and JF) independently reviewed every abstract to determine whether the main topic is about CPV – each was then labeled as include, exclude, or full text needed. The agreed ones, either to include or exclude, were deemed final decision. For those disagreed or marked as full text needed, a second-round screening was performed with another independent reviewer (SM) adjudicating the unresolved ones. All three annotators approved the final set of articles to include for thematic analysis. A coding scheme was derived that covered aspects of interest to the informatics community. The extracted attributes and themes were refined through discussion until a common consensus was achieved.

Results
A summary of the article selection workflow is shown in Figure 1. The integrated database search obtained 57 distinct articles, and we collected another 9 articles from isolated browsing or referral. The 66 articles (57+9) went through first-round screening based on abstract review, where the two annotators agreed to exclude 22 articles as off-topic. Of the 44 articles that got to the second-round screening: 28 had been agreed to include (hence this set was not discussed further), 9 had been disagreed based on abstract, and 7 had been marked as requiring full text review. Of the 16 articles (9+7) discussed among all three annotators, 4 were deemed off-topic and excluded. In the end, 40 articles (44-4) were retained for the synthesis.

Among the final 40 included studies, 10 (25%) dealt with multiple clinical conditions and 6 (15%) did not specify the clinical condition. Breast cancer was the top specific condition (5 studies). All the remaining conditions were covered by only one or two studies, including pressure ulcer, type 2 diabetes, heart failure, hypertension, and soft tissue sarcoma. The frequencies of the involved clinical tasks are summarized in Figure 2a, indicating that treatment (e.g., prescription patterns) was the most studied task with CPV. In Figure 2b we can see that 37 out of the 40 informatics
studies dealt with detection of CPV, which represents the basic task of revealing that variation exists. Characterizing and explaining of CPV follow in counts (27 and 11 studies respectively), where characterizing is identification of associated contexts versus explaining is to establish causal relation with specific variables. Only 5 studies attempted to prevent CPV, and 4 studies tried to generate knowledge by exploiting those detected variations. The derived themes and their sub-themes are summarized in Table 1. The top three by frequency are: Conceptual modeling or methodology design→Algorithm development (6 articles), CDSS implementation and evaluation→Implementation trial (5 articles), and Retrospective auditing of compliance to CPG→machine-assisted chart review (5 articles).

Table 1. The derived themes and sub-themes of informatics research in CPV. CDSS=clinical decision-support system, CPG=clinical practice guidelines, HL7=Health Level Seven.

<table>
<thead>
<tr>
<th>Theme</th>
<th>Sub-themes</th>
<th># of articles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conceptual modeling or methodology design</td>
<td>Algorithm development</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>HL7 message analysis</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Clarity of CPG language use</td>
<td>1</td>
</tr>
<tr>
<td>CDSS implementation and evaluation</td>
<td>Action research</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Analysis of adherence to recommendations</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Implementation trial</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>User satisfaction analysis</td>
<td>1</td>
</tr>
<tr>
<td>Characterizing practice patterns</td>
<td>Clinical documentation</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Field study</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Knowledge discovery by machine learning</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Simulated patients</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Specialty bias</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Telemedicine</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Visual analytics</td>
<td>2</td>
</tr>
<tr>
<td>Literature review and synthesis</td>
<td>Reasons for intended practice deviation</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Advice for practice variation research in informatics</td>
<td>1</td>
</tr>
<tr>
<td>Retrospective auditing of compliance to CPG</td>
<td>Machine-assisted chart review</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Factors associated with CPG noncompliance</td>
<td>3</td>
</tr>
</tbody>
</table>

Discussion
Our scoping review found most informatics studies on CPV have focused around treatment variation and simply the detection of CPV. We see a critical missing piece is to capture patient outcomes in the loop and establish causal relations with CPV, which will enable answering the questions of “why” and “how” certain CPV leads to differential quality in care. Accordingly, more research is needed toward the explaining and learning from CPV.

References
Predicting Hospitalization of COVID-19 Positive Patients Using Machine Learning Methods

Wenyu Song, PhD1,3, Linying Zhang, MS4, Michael Sainlaine, MS1, Mehran Karvar, MD2,3, Min-Jeoung Kang, RN, PhD5, Avery Pullman, BS1, Anthony Massaro, MD1,3, Namrata Patil, MD2,3, Ravi Jasuja, PhD1,3, Patricia C. Dykes, RN, PhD1,3

1Department of Medicine, Brigham & Women’s Hospital, 2Department of Surgery, Brigham & Women’s Hospital, 3Harvard Medical School, Boston, MA, 4Department of Biomedical Informatics, Columbia University, New York, NY, 5College of Nursing, The Catholic University of Korea, Seoul, South Korea

Introduction
Coronavirus disease-19 (COVID-19) pandemic has had a significant impact on the healthcare system globally(1). With millions of infected people, hospitals are facing big challenges due to the lack of beds and other related clinical resources, such as ventilatory support(2). Among COVID-19 infected patients, there is a disproportionately high mortality burden on older adults. Recent studies demonstrate the elderly are most susceptible to severe respiratory illness with increased mortality(3). In the current study, we applied statistical and machine learning techniques to develop personalized risk models using Electronic Health Record (EHR) data. The outcome of the model is whether elderly COVID-19 test-positive patients are hospitalized during the defined time window close to the COVID-19 test date. The long-term goal of this study is to develop an accurate prediction system to facilitate timely hospital care for COVID-19 patients.

Methods
We identified risk factors for COVID-19 using an iterative combination of literature review, qualitative methods (interviews with clinical experts, physicians, who had experience in treating COVID-19 patients) and EHR data exploration (data clinical review and feature engineering). The inclusion and exclusion criteria were developed based on expert opinion and literature review.

We used clinical databases within the Mass General Brigham (MGB) healthcare system, which is a centralized clinical data warehouse for all types of clinical information from multiple Harvard-affiliated hospitals. Available data items include patient demographics, diagnoses, procedures, medication, laboratories, inpatient and outpatient encounter information, and provider data. Elderly patients (age of 65 and above) with at least one positive COVID-19 test at an affiliated MGB site were included in the study. Patients’ hospitalization records within a four-week time window of test date were extracted as the model predictive outcome. Multiple patients’ features, including demographics (age, gender), lab tests (such as albumin) and chronic diseases (such as respiratory disease and heart failure) were obtained from the EHR database as the predictors. Four statistical and machine learning prediction models were trained with these features, including logistic regression, support vector machine (SVM), random forest, and neural network. The precision, sensitivity, specificity, and area under the receiver operating characteristic curve (AUROC) from the test set were used to evaluate the performance of each model. A five-fold cross-validation was conducted within the training set to select hyperparameters that optimize the model performance. The data pre-processing was conducted using R (version 3.3.3). The algorithm training and evaluation were conducted using Python (version 3.7.3).

Results
Using MGB database, we identified 6,334 COVID-19 test-positive elderly patients. By removing patients with high missing values and low data quality, 1,495 patients remained in the final study cohort. In this cohort, the case group (n=459) has hospitalization record (with hospital stay for more than 24 hours after admission) during a four-week window of the COVID-19 test date and the control group (n=1,036) does not have the hospitalization record. Four predictive models were developed using a subset of 29 potential predictive variables (Table 1), including demographics, disease diagnoses and lab values (Table 2). Random Forest and Support Vector Machine achieved the best predictive performance, with an AUROC=0.84.

Among the statistically significant predictors from the logistic regression model, albumin, a plasma protein which is an important index for nutritional status, was found to have the most impact on the model outcome.

Discussion and Conclusion
The COVID-19 pandemic has caused a huge burden to the healthcare system, leading to maximum hospital utilization in many regions of world. To optimize the usage of the clinical resources, it is important for health care providers to identify those at high risk who need timely in-patient services among COVID-19 positive patients. An EHR based accurate predictive tool can facilitate this decision-making process in large populations. There have been studies to
predict patients at high risk of progression and poor outcomes, especially for ICU admission(4, 5). Our goal is to predict all types of hospitalization, which can be used to identify high risk patients in a more general population. We used a multi-hospital study cohort and iterative feature engineering process for model development. The model was validated by using five-fold cross-validation method. Our study showed a robust predictive performance with both tree-based and support vector machine models. Our results also suggested that albumin could be a strong predictor (protector) of hospitalization risk, which is consistent with a previous study(6). Since all our input variables are routinely available patients' features, we are expecting the algorithm can be easily applied in other health care systems in the current and potential future pandemic.

**Table 1.** Summary of EHR-derived measures for model outcomes and predictor variables

<table>
<thead>
<tr>
<th>Definition</th>
<th>EHR Measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hospitalization</td>
<td>14 days prior to COVID-19 test &amp; 14 days post COVID-19 test</td>
</tr>
<tr>
<td>Demographic</td>
<td>Age, Gender</td>
</tr>
<tr>
<td>Vital Signs/Assessments</td>
<td>BMI, Smoking Status, SpO2, Temperature</td>
</tr>
<tr>
<td>Diagnoses</td>
<td>Diabetes, Alzheimer Disease, Cancer, Cardiomyopathy, Cerebrovascular Disease, Chronic Kidney Disease, Chronic Respiratory Disease, Coronary Artery Disease, Cystic Fibrosis, Dementia, Dyslipidemia, Heart Failure, HIV/AIDS, Hypertensive Disease, Immunodeficiency, Liver Disease, Metastatic Solid Tumor, Sickle Cell Disease, Solid Organ Transplant</td>
</tr>
<tr>
<td>Lab Values</td>
<td>Albumin, White Blood Count, Blood Urea Nitrogen, Lymphocyte Count</td>
</tr>
</tbody>
</table>

**Table 2.** Summary of model performance and top features from the logistic regression model

<table>
<thead>
<tr>
<th>Model</th>
<th>AUC (SD)</th>
<th>Precision (SD)</th>
<th>Specificity (SD)</th>
<th>Sensitivity (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic Regression</td>
<td>0.83(0.02)</td>
<td>0.73(0.06)</td>
<td>0.79(0.07)</td>
<td>0.76(0.06)</td>
</tr>
<tr>
<td>Support Vector Machine</td>
<td>0.84(0.02)</td>
<td>0.74(0.05)</td>
<td>0.78(0.07)</td>
<td>0.77(0.08)</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.84(0.02)</td>
<td>0.72(0.07)</td>
<td>0.78(0.07)</td>
<td>0.77(0.07)</td>
</tr>
<tr>
<td>Neural Network</td>
<td>0.83(0.02)</td>
<td>0.70(0.05)</td>
<td>0.80(0.06)</td>
<td>0.75(0.06)</td>
</tr>
</tbody>
</table>

**Top Significant Predictors (based on Logistic Regression)**

- Albumin (Standardized coefficient = -0.101)
- SpO2 (-0.47)
- Temperature (-0.27)
- Cancer (-0.19)
- Blood Urea Nitrogen (0.15)
- Cystic Fibrosis (0.15)
- HIV/AIDS (0.10)
- Solid Organ Transplant (0.07)
- Diabetes (0.07)
- Metastatic Solid Tumor (-0.06)
- Cardiomyopathy (-0.05)
- White Blood Count (0.05)
- Dementia (-0.05)
- Age (0.04)
- Smoking Status (0.04)
- BMI (0.03)
- Heart Failure (-0.03)
- Hypertensive Disease (0.02)
- Coronary Artery Disease (0.03)
- Dyslipidemia (0.02)
- Immunodeficiency (0.02)
- Liver Disease (-0.02)
- Chronic Kidney Disease (0.01)
- Cerebrovascular Disease (-0.01)
- Sickle cell disease/thalassemia (0.01)

**Acknowledgements:** This study was funded by the National Institute on Aging (NIA), 3P30AG031679-10S3, supplement Assisted Living Communities: Transforming Predictive Data into Proactive Care for COVID-19.

**References**

Harmonization of Measurement Codes for Concept-Oriented Lab Data Retrieval

Matthew Spotnitz, M.D., M.P.H.¹, Jason Patterson, B.S.¹, Vojtech Huser, M.D., Ph.D.², Chunhua Weng, Ph.D.¹, Karthik Natarajan, Ph.D.¹
¹Columbia University Medical Center Department of Biomedical Informatics
²National Library of Medicine, Bethesda, MD, USA

Abstract
Measurement concepts are essential to observational healthcare research. We developed five methods for making measurement concept sets by a combination of automated, semi-automated and manual approaches. We validated our concept sets by calculating their frequency in the Columbia University Irving Medical Center (CUIMC) database. We matched basic metabolic panel, lumbar puncture panel and coagulation panel concept sets to the most effective concept set generation method. We intend to expand this work into a multisite study.

Introduction
Laboratory measurement data are essential to observational healthcare research. Prior research efforts have focused on a symbolic representation of laboratory measurements into concept-oriented repositories such as the Medical Entities Dictionary (MED) primarily for clinical care purposes. Today, Logical Observation Identifiers Names and Codes (LOINC) or Systematic Nomenclature of Medicine Clinical Terms (SNOMED-CT) terminologies are commonly used to represent measurement concepts within electronic health record systems. Both LOINC and SNOMED have encoded clinical concepts and have categorized them into hierarchies, which can improve research analyses. However, those hierarchies may include a heterogeneous group of concepts that may not be meaningful for clinical use. A lack of measurement concept harmonization is a challenge for using measurement data in large scale observational research. The issues with measurement concept harmonization are more pronounced in research networks that use healthcare data from multiple sites, such as PCORnet, All of Us (AoU) and the National COVID Cohort Collaborative (N3C). We aimed to achieve two goals with measurement concept harmonization. First, we intended to group concepts by a common biological or chemical assay. Second, we intended to validate the groups of lab tests based on their clinical use.

Methods
Data Source
The Columbia University Irving Medical Center (CUIMC) data warehouse contains lab data from the early 1980s to present day. We converted our data to the Observational Medical Outcomes Partnership Common Data Model (OMOP CDM) and analyzed them in that format.

Algorithms
We uploaded the LOINC table file, which we acquired from the official LOINC website (https://loinc.org/), into a Pandas dataframe. We used semi-automated, automated and manual algorithms to organize the LOINC terms into concept sets (a.k.a. “value sets”). Our algorithms used a combination of medical knowledge and heuristics. For our semi-automated and automated algorithms we used a combination of the following filters: i) component, which is the biological or chemical entity that is measured (i.e. cholesterol) ii) system, which is the kind of specimen from which the sample is drawn (i.e. serum/plasma, blood, urine, etc.) iii) time aspect, which describes the time interval over which the measurement is run (i.e. point measurement, 1 hour measurement, etc.) iv) status, which differentiates active codes that continue to be used in billing data from passive codes that were used historically v) property, which describes the kind of units reported from the measurement (i.e. mass concentration, substance concentration, etc.) vi) analyte, which is the standardized subpart of component that is independent of suffixes and vii) analyte core, which is a standardized subpart of component that is independent of suffixes or precoordinated ratios.

We made a total of five algorithms that included automated, semi-automated and manual approaches. In our first algorithm, we filtered the data frame by four parameters: component, system, time aspect and active use. In a second algorithm, we additionally restricted by property. In a third algorithm, we substituted analyte for component. In a fourth algorithm, we substituted analyte core for component. In a fifth algorithm, we iteratively created concept sets manually in ATLAS, a web-based research platform created by the Observational Health Data Sciences and Informatics.
To validate our concept sets, we calculated their frequency antecedent to cohorts of interest. For example, we calculated the proportions of heart transplant patients who had basic metabolic panel concepts ordered within 30 days antecedent to the operation.

Results

A representative output for our first algorithm, which filtered on component, system, time aspect and active use, is shown in Table 1. The concept set for sodium includes distinct concepts for measurements drawn from blood and from serum/plasma. Our subjective assessment of which algorithms are most effective for making different kinds of concept sets is summarized in Table 2.

Table 1: Sample concept set for sodium in serum, blood or plasma. A total of 4 parameters were used in this algorithm (Component, Time Aspect, System, Active). LOINC NUM, LOINC number; TIME ASPCT, Time Aspect; Pt, Point measurement; mmol/L, millimoles per liter; umol/L, micromoles per liter; mg/dL, milligrams per deciliter; CS ID, Concept Set Identifier.

<table>
<thead>
<tr>
<th>LOINC NUM</th>
<th>COMPONENT</th>
<th>TIME ASPCT</th>
<th>SYSTEM</th>
<th>COMMON NAME</th>
<th>UNITS</th>
<th>CS ID</th>
<th>DESCRIPTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>2947-0</td>
<td>Sodium</td>
<td>Pt</td>
<td>Bld</td>
<td>Sodium</td>
<td>mmol/L</td>
<td>6</td>
<td>Sodium in Serum, Plasma or Blood (point measurement)</td>
</tr>
<tr>
<td>2951-2</td>
<td>Sodium</td>
<td>Pt</td>
<td>Ser/Plas</td>
<td>Sodium</td>
<td>mmol/L</td>
<td>6</td>
<td>Sodium in Serum, Plasma or Blood (point measurement)</td>
</tr>
</tbody>
</table>

Table 2: Subjective assessment of which algorithms are most effective for different laboratory panels.

<table>
<thead>
<tr>
<th>Terms</th>
<th>Method</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heterogeneous</td>
<td>Semi-automated</td>
<td>Basic Metabolic Panel</td>
</tr>
<tr>
<td>Homogeneous</td>
<td>Automated</td>
<td>Lumbar Puncture</td>
</tr>
<tr>
<td>Non-intuitive</td>
<td>Manual</td>
<td>Coagulation Panel</td>
</tr>
</tbody>
</table>

Discussion

We created measurement that were improvements over groupings by the LOINC and SNOMED-CT hierarchies in that they were more meaningful for use in observational research. We included similar concepts from different systems, synchronized concepts by time interval of collection and restricted to codes that were used actively. We grouped our concepts into the panels that are ordered by clinicians in order to identify inaccurate concepts since the components of the panel should occur in similar frequencies.

We used a combination of semi-automated, automated and manual methods to create measurement concept sets. Semi-automated algorithms were most effective for making concept sets from heterogeneously named concepts, such as electrolyte concepts (i.e. Bicarbonate in Serum, Plasma or Blood). Automated algorithms were best for making concept sets with homogeneously named concepts, such as lumbar puncture concepts. Our analysis used data from a single medical center. The maintenance of codes over time is an issue that warrants further investigation. We intend to replicate this analysis on a multiple sites of the OHDSI network.

Conclusion

We created measurement concept sets that are meaningful for use in observational healthcare research using a hybrid of semi-automated, automated and manual methods.

References

MyEDCare: Evaluation of a Smartphone-based Emergency Department Discharge Process

Peter A D Steel, MA, MBBS; David Bodnar, MD; Maryellen Bonito, PA; Jane Torres-Lavoro, MPH; Dona Bou Eid, MHA; Andrew Jacobowitz, PA; Yifan Liu, MS; Amos Shemesh, MD; Robert Tanouye, MD; Patrick Rumble; Daniel DiCello; Rahul Sharma, MD, MBA; Brenna Farmer, MD, MBA; Sandra Pomerantz, MBA; Yiye Zhang, PhD, MS

1NewYork-Presbyterian Hospital Weill Cornell Medical Center, New York, NY; 2Weill Cornell Medicine, New York, NY

Introduction

Poor comprehension of and low compliance with post-emergency departments (ED) care plans increases the risk of unscheduled ED return visits and adverse outcomes. Despite the growth of personal health records (PHR) to support transitions of care, technological innovation focused on the ED discharge process has been limited. There is variable patient engagement with PHRs, including large racial/ethnic disparities. (1) Furthermore, not every patient visiting the ED is part of the associated healthcare system, reducing the incentive for portal use. Recent literature suggests digital communication incorporated into post-ED care can improve patient satisfaction and care quality. We evaluated the feasibility of utilizing MyEDCare, a text message and smartphone-based electronic ED discharge process at two urban EDs piloted in 2019 at NewYork-Presbyterian Hospital Weill Cornell Medical Center and Lower Manhattan Hospital.

Method

MyEDCare sends text messages to adult patients’ smartphones at the time of discharge, containing a hyperlink to a Health Insurance Portability and Accountability Act (HIPAA)-compliant website. Content includes information on therapeutics, new medications, outpatient care scheduling, return precautions, as well as results of laboratory and radiological diagnostic testing performed in the ED. The generic MyEDCare platform display was in English only, but patient-specific discharge instruction content entered by the ED provider and automatically extracted from the EHR was not limited to a single language. Patients discharged from the ED who did not opt-out from MyEDCare were enrolled in the process. The program also excluded patients discharged to locations that would require paper documentation, patients whose employers requested discharge paperwork, patients discharged with a primary psychiatric diagnosis, and patients who were determined by providers to not have sufficient English proficiency. Figure 1 shows the MyEDCare smartphone interface developed by the authors and IT team at the hospital.

Figure 1. MyEDCare smartphone interface

To enroll a patient in MyEDCare, the ED physician confirms eligibility with the patient via verbal consent and verification that he or she is in possession of a functional smartphone with cellular service while in the ED. The physician enters the confirmed phone number into the EHR in a designated section modified specifically for MyEDCare. The physician then places the “ED Discharge Order” in the EHR; this order finalizes MyEDCare enrollment process, triggering the cellular service contracted at our institution to generate the text message. Three text
messages are sent to patients: at time of ED discharge with nurse assistance for initial access of content, as well as 2 and 29 days after ED discharge. The ED nurse guides the patient in their first access of MyEDCare online content on the phone in real-time, confirming content accuracy and patient comprehension, using teach-back method to optimize retention.(2) We collected ED return visit rates, ED staff feedback, and patient satisfaction using ED Consumer Assessment of Healthcare Providers and Systems (ED-CAHPS) patient satisfaction scores.

Results

During the 9-month pilot, 27,713 patients were enrolled in the personal smartphone-based MyEDCare discharge process, accounting for 43% of all treat-and-release patients. A total of 27% of treat-and-release patients further completed MyEDCare ED discharge process, accessing the online content. Scenarios leading to an incomplete MyEDCare discharge process included: lack of smartphone PDF viewing software required to view content, smartphone malfunction or loss of power, data plans prohibiting access to MyEDCare online content, limitations to cellular carrier service, or non-functional hyperlinks. On average, MyEDCare platform was accessed twice per patient. We compared the patient demographics and clinical profile of patients enrolled in MyEDCare; discharged via MyEDCare; did not complete the MyEDCare process; and discharged with standard paper discharge instructions. No clinically meaningful differences in Emergency Severity Index (ESI) levels or diagnosis types were observed between patients who completed MyEDCare ED discharge process compared to those discharged with standard paper discharge instructions. Conversely, we observed significant differences in marital statuses, age, preferred language, insurance status and race. Patients who completed MyEDCare process were more likely to be young, married, commercially insured, and racially White than those who did not. Of the patient who completed MyEDCare discharge process, 9051 (32.7%) patients accessed just once. There are 1009 patients who accessed the online content at least 5 times.

Table 1 compares the ED return visit rate in patients who completed, and did not complete, MyEDCare ED discharge process. Patients discharged via MyEDCare had less frequent unscheduled ED returns at 72 hours and 30 days compared with patients with incomplete MyEDCare enrollment. Patients enrolled in MyEDCare also had less frequent unscheduled ED returns at 72 hours and 30 days compared with patients discharged via standard paper workflows (p-value <0.001). ED providers reported increased efficiency applying the new discharge process, particularly for not having to locate a complete set of discharge papers and then physically deliver the documents to the right nurse caring for the patient at the time of discharge. ED-CAHPS scores for MyEDCare patients demonstrated higher than average scores in the following questions related to nursing care: Nurses explain in way you understand; Nurses listen carefully; Nurses treat you with respect. No other differences in ED-CAHPS questions were determined significant.

Table 1. Comparing ED Utilization between patients who completed and did not complete MyEDCare

<table>
<thead>
<tr>
<th></th>
<th>Complete (N,%)</th>
<th>Incomplete (N,%)</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scheduled 72-hr</td>
<td>39 (0.2%)</td>
<td>26 (0.2%)</td>
<td>0.956</td>
</tr>
<tr>
<td>Unscheduled 72-hr return*</td>
<td>526 (3.1%)</td>
<td>406 (3.8%)</td>
<td>0.003**</td>
</tr>
<tr>
<td>Scheduled 30-day return</td>
<td>152 (0.9%)</td>
<td>86 (0.8%)</td>
<td>0.417</td>
</tr>
<tr>
<td>Unscheduled 30-day return*</td>
<td>1586 (9.4%)</td>
<td>1383 (12.8%)</td>
<td>&lt;0.001***</td>
</tr>
</tbody>
</table>

Discussion

MyEDCare was not appropriate for chronically ill patients from long-term care facilities without smartphones, and patients from vulnerable sociodemographic groups such as patient who are illiterate, non-English speaking or cannot afford a smartphone or the associated carrier data charges. Additional functions of the MyEDCare platform, such as a chat feature, may further enhance post-ED transitions of care in some patient groups who are excluded from PHR use. Additional barriers were predominantly technological. During this pilot our EDs did not use an EHR with an interfacing patient portal, but text messaging may be an alternative to or complement PHRs. Text messaging enjoys high consumer interface familiarity and continues to be the most widely adopted and least expensive technological function on mobile phones. Our findings suggest that EDs and urgent care facilities may consider developing a HIPAA-compliant, text message and smartphone-based discharge process to improve patient-centered outcomes.

References

Evaluating Primary Care Provider Work of Managing Asynchronous Messages through Electronic Health Record Access Logs

Bryan D. Steitz, PhD; Adam Wright, PhD
Dept. of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN

Introduction
Electronic health record (EHR)-based asynchronous communications, also known as inbasket messages, are essential to clinical practice. These message communicate numerous aspects of patient care, including visit scheduling, office administration, availability of test results, and general discussion of patient care plans. Despite their utility to managing clinical practice, a high volume of messages and the time necessary to manage one’s inbox has been identified as a major contributor to job dissatisfaction.1 Previous work has identified that primary care providers (PCPs) receive an average of over 400 messages weekly, which is disproportionately high compared to providers in other specialties.2 Message volume, however, presents a simplified view of the wider task of managing these communications. Few studies have quantitatively investigated the articulation work of asynchronous messaging, such as information seeking, to better understand the work of acting on clinical messages. Developing a better understanding of this articulation work supports the opportunity to make EHR improvements both to the information availability and navigation of the messaging workflow. For example, when a provider receives a medication refill request, we could automatically provide the necessary lab values in the message screen and provide a direct link to create an order. In this study, we utilized the breadth of data captured in action-level EHR audit logs to investigate the information seeking work performed by primary care providers to manage asynchronous clinical communications.

Methods
We identified our cohort of PCPs as any physician, nurse practitioner, or physician assistant who independently treated patients in ambulatory general internal medicine or family medicine clinics at Vanderbilt University Medical Center (VUMC) who had scheduled patient appointments during the month of January, 2020. We extracted from VUMC’s Epic EHR all inbasket messages that a PCP either sent or viewed between January 1, 2020 and January 31, 2020. Messaging data contained a unique message identifier, patient identifier about whom the message was sent, the type of message (i.e. patient message, request for chart review, etc.), the respective message action (i.e. send, read), a unique employee identifier, and a timestamp corresponding to each message event. We grouped message types into six functional categories: patient message, staff message, request for action, result notification, workflow notification, and other. We additionally extracted all EHR access events for each PCP in our messaging cohort. The EHR access data contained an action timestamp, a patient identifier, an employee identifier, and the activity that was accessed (i.e. load schedule, view health maintenance activity, view clinical notes, etc.).

We investigated information seeking work in the context of usage sessions. We defined a usage session as a sequence of activity about a single patient or activity with an undefined patient (i.e. viewing clinic schedule) that occurs within five minutes of any other EHR event about the same patient. We measured information seeking respective to the usage session in which a messaging action was performed. We calculated descriptive messaging and session statistics and calculated information seeking statistics respective to message activity and type.

Results
Between January 1, 2020 and January 31, 2020, there were 1215 PCPs who were involved in 531931 total messages. PCPs in our study accessed messaging functionality in 636486 total sessions, an average of 524 (SD=597) sessions per provider. Messaging sessions lasted an average of 4 minutes and 31 seconds and involved an average of 4.2 different EHR actions. Table 1 presents EHR activity statistics by message activity and type. There were an average of 7.8 and 2.7 EHR actions in sessions that involved sending a message and reviewing a message, respectively. The top EHR actions that were accessed before sending a message and after reading a message are presented in Figure 1.

Discussion
Primary care providers have among the highest rates of burnout across clinical specialties, in part due to EHR use.1 This is among the first studies to investigate the wider scope of acting on these messages by evaluating information seeking relative to messaging actions. Our study was supported by the breadth of data contained within action-level EHR access logs, which allowed us to systematically evaluate EHR activity across an entire organization. By better understanding how EHR activity relates to managing clinical messages, we can begin to devise informatics solutions to improve the efficiency of information seeking behavior. We hypothesize that the burden of messaging work is reduced by the ease with which the EHR enables a response. Future work will seek to develop predictive models, based on message content, to improve information availability and reduce the work information seeking.
Table 1: Session Statistics for Respective to Send and Read Message Actions

<table>
<thead>
<tr>
<th></th>
<th>Patient Message</th>
<th>Staff Message</th>
<th>Request for Action</th>
<th>Result Notification</th>
<th>Workflow Notification</th>
<th>Other</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sent Messages</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total Number of Sent Messages (%)</td>
<td>88,365</td>
<td>22,596</td>
<td>52,530</td>
<td>22,537</td>
<td>26,944</td>
<td>2,050</td>
<td>215,022</td>
</tr>
<tr>
<td>Mean Messages per PCP (SD)</td>
<td>88.2 (113.0)</td>
<td>24.5 (33.6)</td>
<td>45.6 (61.4)</td>
<td>27.2 (47.7)</td>
<td>28.4 (45.1)</td>
<td>4.4 (8.5)</td>
<td>177 (197)</td>
</tr>
<tr>
<td>Total Number of Sessions with Message Send (%)</td>
<td>116,295</td>
<td>29,562</td>
<td>67,729</td>
<td>30,507</td>
<td>32,708</td>
<td>2,545</td>
<td>254,818</td>
</tr>
<tr>
<td>Mean Sessions per PCP (SD)</td>
<td>116.0 (150.0)</td>
<td>32.1 (44.6)</td>
<td>59.9 (82.2)</td>
<td>36.8 (67.6)</td>
<td>34.5 (49.7)</td>
<td>5.5 (11.7)</td>
<td>210 (234)</td>
</tr>
<tr>
<td>Mean Seconds per Session (SD)</td>
<td>117.3</td>
<td>154.2</td>
<td>193.6</td>
<td>104.1</td>
<td>205.7</td>
<td>131.1</td>
<td>140.0</td>
</tr>
<tr>
<td>Mean EHR Events per Session (SD)</td>
<td>5.2 (5.4)</td>
<td>7.8 (8.2)</td>
<td>10.0 (9.9)</td>
<td>5.1 (5.7)</td>
<td>13.2 (10.8)</td>
<td>6.0 (7.8)</td>
<td>7.8 (8.8)</td>
</tr>
<tr>
<td>Mean EHR Events per Session Before Send (SD)</td>
<td>4.8 (4.7)</td>
<td>6.6 (6.1)</td>
<td>7.7 (7.0)</td>
<td>4.9 (4.6)</td>
<td>9.8 (7.6)</td>
<td>5.7 (5.9)</td>
<td>7.1 (6.6)</td>
</tr>
<tr>
<td><strong>Read Messages</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total Number of Read Messages (%)</td>
<td>96,660</td>
<td>21,135</td>
<td>125,258</td>
<td>79,949</td>
<td>31,827</td>
<td>4,571</td>
<td>359,400</td>
</tr>
<tr>
<td>Mean Messages per PCP (SD)</td>
<td>97.5</td>
<td>24.2</td>
<td>118.0</td>
<td>81.0</td>
<td>37.6</td>
<td>9.0 (1.3)</td>
<td>321.0</td>
</tr>
<tr>
<td>Mean EHR Events per Session (SD)</td>
<td>5.5 (3.1)</td>
<td>118.0</td>
<td>81.0</td>
<td>37.6</td>
<td>9.0 (1.3)</td>
<td>321.0</td>
<td></td>
</tr>
<tr>
<td>Total Number of Sessions with Message Send (%)</td>
<td>128,332</td>
<td>27,651</td>
<td>174,859</td>
<td>148,426</td>
<td>38,493</td>
<td>5,322</td>
<td>498,365</td>
</tr>
<tr>
<td>Mean Sessions per PCP (SD)</td>
<td>128.0</td>
<td>29.5</td>
<td>164.0</td>
<td>148.0</td>
<td>42.2</td>
<td>10.4</td>
<td>436.0</td>
</tr>
<tr>
<td>Mean Seconds per Session (SD)</td>
<td>55.1 (101.0)</td>
<td>49.3 (103.0)</td>
<td>33.2 (91.3)</td>
<td>39.0 (96.8)</td>
<td>22.0 (66.9)</td>
<td>24.5 (78.2)</td>
<td>40.0 (94.7)</td>
</tr>
<tr>
<td>Mean EHR Events per Session (SD)</td>
<td>2.3 (2.6)</td>
<td>2.9 (3.4)</td>
<td>3.4 (4.5)</td>
<td>2.4 (3.4)</td>
<td>2.6 (3.0)</td>
<td>3.1 (5.3)</td>
<td>2.7 (3.3)</td>
</tr>
<tr>
<td>Mean EHR Events per Session After Read (SD)</td>
<td>2.2 (2.5)</td>
<td>2.8 (3.2)</td>
<td>3.2 (4.3)</td>
<td>2.1 (2.7)</td>
<td>2.5 (2.7)</td>
<td>2.7 (4.3)</td>
<td>2.5 (2.7)</td>
</tr>
</tbody>
</table>

Figure 1: Top Five Pre-Send and Post-Read EHR Actions Measured as the Percent of Total Sessions.

References
FHIRT ime: Standardizing Temporal Patterns Identified from Clinical Narratives Using HL7 FHIR

Daniel Stone, BS¹, Sijia Liu, PhD¹, Yuan Luo, PhD², Wen Andrew, MS¹, Nansu Zong, PhD³, Luke V. Rasmussen, MS², Prakash Adekanattu, PhD³, Pascal S. Brandt, MSc⁴, Jennifer A. Pacheco, MS², Fei Wang, PhD³, Cui Tao, PhD⁵, Jyotishman Pathak, PhD³, Hongfang Liu, PhD¹, Guoqian Jiang, MD, PhD¹

¹Mayo Clinic, Rochester, MN; ²Northwestern University, Chicago, IL; ³Weill Cornell Medicine, New York, NY; ⁴University of Washington, Seattle, WA. ⁵The University of Texas Health Science Center at Houston, Houston, TX

Introduction
A temporal dimension is essential for the interpretation of clinical narratives. Lack of standardized temporal representation hinders the portability of clinical phenotyping applications using electronic health record (EHR) data. Existing standardized clinical models including HL7 Fast Healthcare Interoperability Resources (FHIR) define data types that can be used to specify the complex timing of events and processes that occur in clinical applications. However, these time models are meant for structured data, and remain untapped with unstructured and semi-structured clinical data. In a previous study, we developed a clinical data normalization pipeline known as NLP2FHIR, which standardizes and integrates unstructured and structured electronic health record (EHR) data using FHIR. The objective of this study is to develop an approach known as FHIRT ime to demonstrate the capability of FHIR for representing temporal patterns identified from clinical narratives. We leveraged the Time Event Ontology (TEO) for standardizing time patterns. Two temporal corpora were used for the preliminary evaluation of the FHIRT ime approach.

Materials and Methods
Materials: TEO is an ontology designed to support semantic representation of, and reasoning about complex temporal relations of clinical events. THYME Corpus is a human-annotated corpus of medical records in which the temporal relations between different events, occurrences, states, dates, and procedures are clearly annotated. Its annotations are organized as Event, TIMEX3 (i.e., phrases that contain time information, having classes of Date, Time, Duration, Quantifier, etc.), TLINK (i.e., relation annotations with subtypes of before, contains, overlap, begins-on, and ends-on), and other types. The 2012 i2b2 Challenge Corpus contains 310 discharge summaries annotated for temporal information (i.e., clinically significant Event, TIMEX3, and TLINK). In simpler terms, TLINK are used to describe a temporal relation from one entity (Event or TIMEX3) to another entity or Document or Section Time. One might view these relations as graph-like with entities serving as vertices and TLINK serving as directed or undirected edges.

Methods: We first analyzed the FHIR built-in time artifacts, and the temporal annotations from two temporal corpora (i2b2 Challenge and THYME) and aligned them with high-level time classes defined in TEO. Second, we developed an approach (called “FHIRT ime”) that handles temporal patterns utilizing FHIR built-in time artifacts and its extension mechanisms. Specifically, the Event annotations can be a clinical event or a non-clinical event. If an Event annotation can be represented as an instance of a FHIR resource such as Condition, Procedure, or Medication Statement, it was treated as a clinical event; otherwise, it was treated as a non-clinical event that was then represented as an instance of the FHIR Observation resource. For the TIMEX3 annotations, if they are associated with an Event, we used the FHIR built-in time artifacts (e.g., fhir:Condition.onset[x]) to represent them. If they have a temporal relationship with an Event, we treated them as a TLINK. For the TLINK annotations, we created a collection of FHIR extensions leveraging TEO time relationships and used them to handle TLINK relations between Events, Times or Event-Time. Third, we evaluated the FHIRT ime approach by converting the temporal annotations from the two temporal corpora into the FHIR-based representation. Specifically, we invoked the NLP2FHIR to identify clinical events with standardized codes (e.g., SNOMED CT or RxNorm). We then mapped the identified clinical events with the Event annotations in the two corpora. We analyzed the coverage of converted temporal patterns in FHIR to demonstrate the robustness of proposed FHIRT ime approach.

Results
Figure 1 shows the alignment results between FHIR, TEO and temporal annotations. Five core FHIR clinical resources Condition, Procedure, MedicationStatement, Observation and FamilyMemberHistory are represented as subclasses of teo:Event. Ten FHIR datatypes are represented as the subclasses of teo:Time, teo:TimeInstance or teo:TimeInterval, and four of them (fhir:date, fhir:time, fhir:Timing, and fhir:Duration) are aligned with TIMEX3
subtypes. We created five FHIR extensions for five TEO time relationships (teo:before, teo:contain, teo:overlap, teo:start, and teo:finish) that aligned with TLINK types.

We were able to convert temporal annotations from two corpora into the FHIR-based representations. For TIMEX3 annotations, we were able to fully represent TIMEX3:DATE from the i2b2 dataset as it represented dates in ISO format, whereas only 4% of THYME TIMEX3:DATE could be represented as fhir:date without any parsing; this improved to 65% when using the temporal NLP Java library SUTime (Stanford University) for parsing. THYME TIMEX3:DATE representations pose a challenge as even annotated TIMEX3:DATE have a loose semantic definition of date including relative dates ("yesterday") as well as phrases that refer to a date ("at that time"). For other TIMEX3 subtypes such as TIME (e.g. "16:43"), DURATION (e.g. "In the interim"), and QUANTIFIER (e.g. "multiple times"), fhir:string was used to represent them due to most annotations of these subtypes consisting of time phrases. For TLINK annotations, the FHIR extensions based on TEO relationships were able to fully represent all TLINK subtypes from the two datasets.

Discussion
In this study, we explored the capability of FHIR to represent temporal patterns identified from clinical narratives and we found that FHIR built-in artifacts can well cover clinical events and TIMEX3 subtypes. For the TLINK annotations, we were able to leverage the FHIR extension mechanism, in which the TEO time relationships were leveraged. THYME captures TIMEX3 from a linguistic approach when describing temporal relations from an NLP grammatical structure perspective, but this poses significant challenges when attempting to resolve these phrases to a standardized date as required by fhir:date. While using NLP to capture temporal relations is a challenging task, temporal relations in unstructured text can capture a clinician’s interpretations of causality that can supplement structured data in EHR systems. Explainable artificial intelligence applications could potentially leverage this to provide more clinician-friendly explanations through causality. There are a number of limitations in this study. First, we focused on common temporal patterns (i.e., Event, TIMEX3 and TLINK). We noticed that THYME corpus also contains some ALINK (aspectual link) annotations which need to be investigated in a future study. Second, we did not address attributes of Event annotations (e.g., polarity, degree, contextual modality) in this study, but we have defined a collection of FHIR extensions in NLP2FHIR to handle many of these NLP-specific data elements. Third, we noticed that many of the TIMEX3 annotations with the type of Date or Time were a time phrase rather than an exact date or time, indicating the complexity of temporal patterns from clinical narratives. For the FHIR representation, we used fhir:string instead of fhir:date or fhir:time to represent such time phrases.

As deep phenotyping is shifting the focus from case ascertainment or cohort identification to precise and comprehensive characterization of a phenotype, we consider the outcome of this FHIRTime work would enable a better characterization of the temporal trajectory of a phenotype in a standardized and reproducible manner. Future work will be focused on integrating the FHIRTime with the NLP2FHIR pipeline to enhance its temporal modeling and temporal reasoning capability.

Acknowledgment
Research reported in this publication was supported by National Institutes of Health under the awards FHIRCat (R56EB028101), PhEMA (R01 GM105688), TEO (R01 LM011829, R01 AI130460).

References

Figure 1. The alignment results among FHIR built-in time artifacts (Red), and the temporal annotations from two temporal corpora (Blue) and the high-level TEO classes (Black)
PASC CKD: revealing the progression trajectories of sustained COVID-19-related renal injury using real-world evidence

Jing Su1, Tian He1, Michael Eadon2, Xiaochun Li1, Stanley Taylor1, Zuoyi Zhang1, pengyue Zhang1, Chi-Mai Nguyen3, Mary Lynn Davis-Ajami3, Tae-Hwi Schwantes-An4, travis johnson1, Kun Huang1, Zhaorui Liu5, Ziyang Tang5, Baijian Yang5, Qianqian Song6,7
1Department for Biostatistics and Health Data Science, 2Department of Medicine, 3School of Nursing, 4Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, IN 46202, USA; 5Department of Computer and Information Technology, Purdue University, West Lafayette, IN 47907, USA; 6Center for Cancer Genomics and Precision Oncology, Wake Forest Baptist Comprehensive Cancer Center; 7Department of Cancer Biology, Wake Forest School of Medicine, Winston Salem, NC, 27157, USA;

Introduction

The COVID-19 pandemics has permanently changed the health projections of those who survived and cast shadows over the health of 10 ~ 20% world population, known as past acute sequelae SARS-CoV-2 infection (PASC). A major PASC clinical condition is Sustained COVID-19-related Renal Injury (Sus-CovRI). However, the middle-to-long term health impact of such COVID-induced or exacerbated chronic kidney disease (CKD) is largely unknown. In this work, we use the longitudinal Optum Clinformatics™ claim data and group-based trajectory model (GBTM)5 to explore the middle-term progression patterns of Sus-CovRI.

Methods and Results

Cohort characteristics. As shown in the CONSORT flow diagram (Fig. 1), in this study, we use the Optum Clinformatics™ Data Mart (Jan 1, 2007 to September 30, 2021) of claims data. The study cohort consists of about 19 million patients that have been enrolled between 01/01/2020 and 09/30/2021. Within this cohort, 279,071 (1.47%) patients were diagnosed with COVID-19 (ICD10:U07.1). The demographic and clinical characteristics of the overall study cohort and COVID-19 positive sub-cohort are show in Table 1. For the overall study cohort, the demographic pattern is representative of the U.S. population, with 10.5% as African Americans, 13.8% as Hispanic Americans, and 51.5% as females. For the COVID-19 positive sub-cohort, it recaptures known demographical features as reported by peers, which has significantly increased infection risks for African Americans and Hispanics, with odds ratios of 1.53 (95%CI: 1.51, 1.54; P-value < 2.2e-16) and 1.73 (95%CI: 1.71, 1.75; P-value < 2.2e-16), respectively.

The Sus-CovRI Phenomics. A Sus-CovRI phenomics is defined as a comprehensively and systematically curated collection of phenotypic information for biomedical informatics data analysis of PASC CKD. We extend our previously developed Common CLINic Index for Chronic diseases (CLINIC) CDM to OMOP CDM and comprehensively outline the general health conditions during the development and progression of common chronic diseases. The selection of these clinical indices is a tradeoff of the following considerations: 1) reflect health status and risks of chronic disease with aging; 2) clinically measured and provide strong evidence; 3) available in Optum; 4) relevant to COVID-19 syndromes; and 5) ready to use.

Longitudinal summarization of RWE data. The longitudinal RWE data is grouped into the following two time windows (Fig. 2): 1) the COVID-19/PASC window (red/blue shaded window) is defined as the time range from 14 days before the 1st positive test to the last follow-up; and 2) the Pre-COVID window (green shaded) is defined as
starting from 30 days before the 1st positive test and retrospectively tracked for 2 years. Transient features such as HbA1c levels in the Pre- or COVID-19 period are aggregated or summarized. Sustained conditions such as diabetes and hypertension will be tracked back to all available data. Transient features such as estimated glomerular filtration rate (eGFR) during COVID-19/PASC period are aggregated by day and used as longitudinal data. Baseline is defined as 14 days before the 1st positive test. Patients’ clinical conditions at baseline are composed of the aggregated or summarized transient features in the Pre-COVID window as well as sustained conditions in and before the Pre-COVID period. Such longitudinal data summarization is able to capture the renal injury events associated with the SARS-CoV-2 infection.

Discovered Sus-CovRI trajectory groups. As shown in Tab. 1, the trajectory group learning cohort (n = 99,678, 35.7% COVID-19 patients) is defined as patients who have positive COVID-19 diagnosis and have at least one eGFR values or serum creatinine measurements in the Pre-COVID window. Within this cohort, a COVID Survivor sub-cohort (n = 5,127, 1.84% of COVID-19 patients) is defined as patient with at least one eGFR value after 90 days post the initial COVID-19 diagnosis. Pre-COVID and Post-COVID CKD progression trajectories are identified using the BGTM on the Optum longitudinal eGFR data derived from serum creatinine laboratory results with CKD-EPI 2009 formula. Three Pre-COVID renal trajectories and their corresponding Post-COVID trajectories are discovered: 1) healthy group (blue curve) without existing renal condition. Among them, 2.3% demonstrated long-lasting renal injury after recovering from COVID; 2) mild CKD group (yellow curve), with 16.7% showed PASC CKD symptom; and 3) severe CKD group, with 48.7% patient showed COVID-exacerbated CKD (red curve).

Conclusion

Our work is one of the first large-scale profiling of the disease progression trajectories of the PASC CKD population. We reveal major CKD progression patterns and pave ways to better understanding the complexity of PASC CKD, determining clinical characteristics and risk factors, and providing RWE-based clinical decision support for managing COVID-19 survivors’ kidney health.

Reference

Gender Differences in Time to Diagnosis through Fairness and Time Variant Evaluation of EHR Data

Tony Y. Sun, MA¹, Jennifer L. Chen, BA¹, Harry Reyes Nieva, MAS¹, Oliver J. Bear Don’t Walk IV, MA¹, Jaan Altosaar, PhD¹, Noémie Elhadad, PhD¹
¹Columbia University, New York, NY, USA

Introduction

Recent research has identified differences in disease presentation, progression, and treatment responses across a number of health conditions among men and women. Here, we focus on time to diagnosis (TTD), traditionally defined as the interval from first alert symptoms to patient disease diagnosis, and explore potential TTD-related gender disparities. Specifically, across a large number of conditions, we quantify differences in time to diagnosis across women and men in a longitudinal fashion.

Methods

Dataset Using data from Columbia University Irving Medical Center, we selected patients with at least 3 years of continuous observation in 2010-2020. The study was approved by the Columbia IRB. Using 121 established phenotype definitions publicly available from the OHDSI HERA study, we constructed 121 corresponding cohorts, where entry time in the cohort corresponds to time of diagnosis. For each patient in a given cohort, we extracted their gender and all coded condition occurrences observed within 3 years prior to phenotype diagnosis. Along with each coded condition, we kept track of the date of its first occurrence in a patient longitudinal record. The final dataset comprises 533,566 women (mean age, 49.5 years) and 412,498 men (mean age, 49.1 years) with 14,900 unique condition codes.

Time-Variant Model Fairness in Diagnosis Classification For each phenotype, we trained a gender-agnostic binary classifier that diagnoses patients based on their clinical observations (i.e., first-occurrence of condition codes) documented up to 3 years prior to the official disease diagnosis. As such, the trained classifier is “fully knowledgeable” as it contains all observations up to time of diagnosis, particularly the relevant symptoms (e.g., abdominal pain indicating Crohn’s disease). We measured the classifiers’ performance on a cohort of patients with decreasing levels of right censoring (Figure 1(a)) up to 3 years prior to their diagnosis. In essence, this testing procedure aims to mimic a provider diagnosing patients at increasing time steps with knowledge limited to patient history accumulated thus far.

In line with established fairness literature, we measure the equality of opportunity of the positive and negative class (recall and specificity gap, respectively), and the precision and accuracy gap at each test window. For example, recall gap is the recall of diagnosing male patients minus the recall of female patients. In diagnosis, recall gap is important as minimizing false negatives ensures patients with a disease are not left untreated. Our approach to analyzing fairness gap over time is extensible to any definition of fairness relevant to the data domain.

In our analysis, we evaluate bias on the test data in a longitudinal manner, decreasing the amount of right-censoring at each time step and measuring the fairness gap over time. To measure the average fairness gap as information accumulates over time, we leverage mean squared error (MSE) and the overall directionality of bias over each time windows, combining these concepts into mean squared discrimination (MSD).

Results

Out of 121 phenotypes, 65 were excluded from the study because of low prevalence in our patient population. The following results pertain to the remaining 56 phenotypes (Figure 2). We use Crohn’s Disease as an illustrative example.

As the amount of patient history increases, the Crohn’s diagnosis model’s recall for both genders increases, but performance for men is consistently higher than women (Figure 1(b)); we see this in the positive, male-favored MSD recall metric for Crohn’s (0.097), which indicates a consistently male-favored odds of diagnosis.

*We recognize gender is a social construct while sex is a biological variable. We use the terms “gender,” “men,” and “women” based on available labels in our datasets.
†https://data.ohdsi.org/HERACharacterization/
Figure 1: (a) Test sets are generated with varying levels of right censoring, mimicking varying amounts of patient history available to providers (b) Gender-stratified recall at various time-bins for the Crohn’s diagnosis classifier.

Discussion and Conclusion

We propose an approach to evaluate gender differences in TTD in a longitudinal fashion. Using recall MSD as a proxy for diagnosis bias, we observe gender bias in all 56 phenotypes in the study, with 36 skewed towards earlier diagnosis for men (e.g. Crohn’s). Furthermore, women had more documented conditions prior to diagnosis (for instance, in the Crohn’s cohort, on average men had 17 conditions documented prior to diagnosis and women had 19). This might indicate that while women present the right conditions earlier to help detect the presence of a condition, they also present many other conditions that may confuse the classifier (and potentially providers as well), thus rendering a male-biased recall. The marked gender differences we find in recall gaps raise open questions about disparities in healthcare.

References

Physicians with hybrid skillsets in digital health and informatics facilitate necessary innovation in mental health care delivery: A physician engagement strategy

Tania Tajirian, MD, CCFP, FCFP, DTMPH\textsuperscript{1,2}, Brian Lo, MHI\textsuperscript{1,2}, Gillian Strudwick, RN, PhD, FAMIA\textsuperscript{1,2}, Lydia Sequeira, MHI\textsuperscript{1,2}, Damian Jankowicz, PhD\textsuperscript{1}

\textsuperscript{1}Centre for Addiction and Mental Health, Toronto, Ontario, Canada; \textsuperscript{2}University of Toronto, Toronto, Ontario, Canada

Introduction

Delivering mental health care in innovative ways that do not solely rely on face-to-face care remains an area of priority as the COVID-19 pandemic exacerbates the mental health of Canadians across the country \textsuperscript{1}. However, the effective implementation and adoption of innovative approaches (which may include digital health) into practice continues to remain a challenge \textsuperscript{2}. In many cases, despite best practices from implementation science and change management (e.g., Consolidated Framework for Implementation Research \textsuperscript{3}), many of these innovations fail to become fully embedded into practice at an organization. More problematically, as outlined in the Clinical Adoption Meta-Model\textsuperscript{4}, innovations that are suboptimally implemented may eventually lead to workarounds and unintended consequences.

With this problem in mind, at our large academic and teaching mental health hospital in Toronto, Ontario, we implemented the Physician Engagement Strategy\textsuperscript{5} to address the ongoing burden related to the use of the electronic health record (EHR) system. In this strategy, physicians have the opportunity to engage in driving change through both virtual and in-person modalities. As we implemented this strategy throughout the pandemic, we began to notice how these initiatives have inspired a new generation of “hybrid physicians”\textsuperscript{6} committed to leading changes across the organization in both digital health and practice-related capacity. The impact of these interventions has led to the development of a unique workforce of physicians that strive to reduce silos across Academic Divisions and foster a sense of excitement and collaboration towards innovations in clinical care. In this podium presentation, we share how the Physician Engagement Strategy has led to the transformation of the organizational and informatics workforce at our institution and discuss lessons learned and recommendations for translating this vision into other organizations.

Methods

The Physician Engagement Strategy\textsuperscript{5} is a multi-pronged strategy that includes multiple interventions that aim to engage physicians, inspire change, and measure impact. The development of this initiative has been driven by an academic divisional tour of the ongoing EHR-related challenges and a baseline survey with front-line physicians across the organization \textsuperscript{7}. Based on these findings, several initiatives that focus on inviting clinicians to be agents of change have been developed and implemented. This included:\textsuperscript{5}

- **Physician Think Tanks** – A cross-divisional physician forum that meets every month with stakeholders from IT leadership (Health Records, Clinical Applications), clinical informatics nurses, clinical leadership, and pharmacy and laboratory informaticians. The main objective of this forum is to foster discussion across the members of different departments and divisions in order to support decision-making and strategy development.

- **‘SWAT’ Teams**\textsuperscript{5} – A team comprising of clinician and IT administrators that work with front-line clinicians from each academic division to identify and address issues related to the EHR in an agile manner. Once an issue or challenge has been identified, it is then triaged based on the type of solution needed and the level of changes required (e.g., simple fix vs. new feature).

- **Physician portal and monthly newsletters** - A portal that is a one virtual stop for physicians to find information about a variety of wellness initiatives, personal stories, and educational presentations such as Grand Rounds. Monthly newsletters to communicate IT changes and EHR improvements that are completed through the think tanks and SWAT tours.
Results

Between 2019 and 2021, two rounds of SWAT and over 24 Physician Think Tank meetings were conducted. This work has collectively inspired an evolving workforce dedicated to advocating and driving change in innovative mental health care as part of the organization. These observations are detailed below.

As part of the SWAT team initiative, physicians had a unique opportunity to become closely engaged with IT and informaticians to report any issues that hinder their overall experience with digital tools. While the original intent was a focus on the use of the EHR system, a range of issues have been reported, including speech recognition technology, virtual care, etc. The reporting of these issues has since led to a broader discussion with other clinical departments (e.g., telemedicine) and practice changes. As such, the SWAT initiative has become a unique, low-burden, channel for physicians to advocate for challenges and opportunities to improve the delivery of mental health care.

In addition to the work from the SWAT initiative, the Physician Think Tank has evolved over time to support the development of ‘hybrid physicians’ in various ways. While the initial iteration of the Physician Think Tank comprised of a single physician liaison from each academic division, the membership has since grown to include professional practice stakeholders and physicians with other responsibilities (e.g., research) interested in informatics. Likewise, the topics brought up by the committee members have grown significantly. Recent topics have included virtual care and broader practice workflows that have an intersection with informatics. As such, members of the Think Tank have become advocates for informatics and mental health care on behalf of the user group they represent.

Discussion

In this work, we share our perspectives and experience on how the implementation of our Physician Engagement Strategy has supported the transformation of the organizational and informatics workforce to become advocates for innovative mental health care. As observed in our SWAT initiative, it has allowed for a decentralized approach to participating in the optimization of our EHR system and digital health infrastructure. Alongside the increased membership in the Think Tank, these opportunities have collectively fostered talent and interest amongst physicians in advocating for digital health with IT and other leadership stakeholders.

Conclusion

Fostering a physician workforce dedicated to informatics and the delivery of innovative mental health care has become instrumental in the implementation success of digital health. In this presentation, we share our reflections and experience with how the Physician Think Tanks and SWAT teams have led to a team of ‘hybrid physicians’ at the intersection of medicine and digital health.

Acknowledgements

The authors would like to acknowledge the organization’s physicians and leadership teams for their support in these initiatives.

References

Urologist Use and Perceptions of the EHR: Perspectives from a Surgical Specialty

Hung-Jui Tan, MD, MSHPM1,2; Arlene Chung, MD, MHA, MMCi2,3; David Gotz, PhD1,4; Allison Deal, MS2; Antonia Bennett, PhD2,5; Matthew Nielsen, MD, MS1,2; and Ethan Basch, MD, MPH2,3,5

1Department of Urology, School of Medicine, University of North Carolina, Chapel Hill, NC
2Lineberger Comprehensive Cancer Center, School of Medicine, University of North Carolina, Chapel Hill, NC
3Department of Medicine, School of Medicine, University of North Carolina, Chapel Hill, NC
4School of Information and Library Science, University of North Carolina, Chapel Hill, NC
5Department of Health Policy & Management, Gillings School of Global Public Health, University of North Carolina, Chapel Hill, NC

Introduction: Electronic health records (EHRs) serve as the backbone of modern care delivery with increasing emphasis given now to improving quality through embedded tools such as clinical decision support. Potential advances, however, may be hampered by persistent usability issues with EHRs, and relatively little attention has been given to surgery and surgical specialties, such as urology. To push forward, we explored the use and perceptions of EHRs among urologists.

Methods: We conducted a national sequential explanatory mixed methods study. Via the 2019 American Urological Association Census, we surveyed practicing urologists on EHR function use and impact on clinical efficiency and patient care. Urologists rated the impact of the EHR on clinical efficiency and patient care using a 5-point Likert scale (1: Strongly Disagree to 5: Strongly Agree). We then calculated the net favorability—the proportion agreeing minus the proportion disagreeing—as a summary measure. We also created a binary measure for agreeing (Strongly Agree and Agree) to identify associated characteristics through bivariable and multivariable analyses. Next, we interviewed 25 respondents on their EHR experiences. Coding-based analysis was applied to identify concepts and themes. These qualitative findings were then integrated with survey findings through a joint display and weaving narrative.

Results: In 2019, 96.4% of practicing urologists used an EHR for a weighted sample of 12,366. Over 90% used the EHR for charting, viewing results, and order entry (Figure). Fewer used the communication functions: 79.6% communicated with providers internally, 64.6% sent/received information externally, and 59.0% exchanged messages with patients. Few utilized information from alerts (43.4%) or reported data to clinical registries (28.4%). In terms of perceptions, less urologists felt the EHR increased clinical efficiency (35.8% agree vs. 47.4% disagree; -11.6%) while more felt it improved patient care (43.1% agree vs. 33.4% disagree: +9.7%). Perceptions differed based on experience, workload, and practice setting (p<0.05), corresponding respectively to themes on prior EHR exposure, administrative burden, and infrastructure/support (Joint Display). Thematically, perceptions of clinical efficiency related to information management (e.g., access vs. excess vs. incoherent/inaccurate data) while the impact on patient care focused on patient safety and the patient-physician relationship (e.g., safeguards vs. extraneous non-clinical tasks).

Figure: Use of Specific EHR Functions
Joint Display: Perceived Benefit of EHR stratified by Key Characteristics and Related Themes

<table>
<thead>
<tr>
<th>Key Characteristics</th>
<th>Perceived Benefit* (Quantitative)</th>
<th>Themes with Representative Quotes (Qualitative)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Clinical Efficiency</td>
<td>Patient Care</td>
</tr>
<tr>
<td>Overall</td>
<td>-11.6%</td>
<td>+9.7%</td>
</tr>
<tr>
<td>Years in Practice</td>
<td></td>
<td></td>
</tr>
<tr>
<td>≤10</td>
<td>+7.5%</td>
<td>+29.9%</td>
</tr>
<tr>
<td>11–20</td>
<td>-11.2%</td>
<td>+12.3%</td>
</tr>
<tr>
<td>21–30</td>
<td>-28.0%</td>
<td>-7.5%</td>
</tr>
<tr>
<td>≥30</td>
<td>-20.8%</td>
<td>-1.8%</td>
</tr>
<tr>
<td>Clinical Work Hours / Week</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;40</td>
<td>-4.3%</td>
<td>+18.1%</td>
</tr>
<tr>
<td>40–49</td>
<td>-10.6%</td>
<td>+10.1%</td>
</tr>
<tr>
<td>50–59</td>
<td>-7.5%</td>
<td>+10.1%</td>
</tr>
<tr>
<td>≥60</td>
<td>-22.3%</td>
<td>+1.2%</td>
</tr>
<tr>
<td>Practice Size</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;5 urologists</td>
<td>-16.4%</td>
<td>+0.3%</td>
</tr>
<tr>
<td>≥5 urologists</td>
<td>-8.6%</td>
<td>+15.5%</td>
</tr>
<tr>
<td>Practice Type</td>
<td>Academic</td>
<td>-9.5%</td>
</tr>
<tr>
<td>Multispecialty</td>
<td>-6.5%</td>
<td>+16.0%</td>
</tr>
<tr>
<td>Public</td>
<td>-2.4%</td>
<td>+20.6%</td>
</tr>
<tr>
<td>Private Hospital</td>
<td>-18.7%</td>
<td>-0.1%</td>
</tr>
<tr>
<td>Urology Group</td>
<td>-19.7%</td>
<td>-5.5%</td>
</tr>
<tr>
<td>Solo Practice</td>
<td>-11.2%</td>
<td>-4.7%</td>
</tr>
<tr>
<td>Rurality</td>
<td>Urban</td>
<td>-9.2%</td>
</tr>
<tr>
<td></td>
<td>Rural</td>
<td>-33.1%</td>
</tr>
</tbody>
</table>

*Percent who “Agree” or “Strongly Agree” minus the percent who “Disagree” or “Strongly Disagree”

**Conclusions:** Though EHRs have become widely adopted, the degree of usage varies, and EHRs continue to engender negative feelings, particularly by busy urologists in smaller, less resourced practices. With respect to clinical efficiency, concerns center on information management. While urologists found that the EHR offered greater access to information, they also found that this information can be inaccurate and superfluous, creating administrative inefficiencies that are magnified by a compressed digital work schedule due to surgery and other procedures. From a patient care standpoint, urologists highlighted the potential of the EHR to help them avoid adverse outcomes with relatively little mention of evidence-based medicine or guideline-concordant care. Urologists also highlighted the potential harms of inaccurate information and the impact on physician-patient interaction. EHR-based tools that ameliorate administrative burden and simplify information management may be best positioned for effective use. Moreover, framing the use case of specific functions and tools on how they can improve outcomes and performance and/or how they respect the physician-patient relationship may further promote dissemination and implementation.

**Acknowledgements**

Hung-Jui Tan, MD, MSHPM was supported by a Mentored Research Scholar Grant in Applied and Clinical Research, MRSG-18-193-01-CPPB, from the American Cancer Society as well as the NIH Loan Repayment Program. These funding sources had no role in the design, conduct, analysis, or decision to publish the manuscript.
ADNet: Identify biomarkers of Alzheimer Disease with MRI and EMR data using Deep Neural Networks

Ziyang Tang, PhD1, Qianqian Song, PhD2, Jing Su, PhD3, Baijian Yang, PhD1
1Department of Computer and Information Technology, Purdue University, West Lafayette, IN, USA; 2Department of Cancer Biology, Wake Forest School of Medicine, Winston Salem, NC, USA; 3Department of Biostatistics, Indiana University School of Medicine, Indianapolis, IN, US

Abstract We develop a data-driven deep learning framework to identify the brain structural differences among the early-stage patients of Alzheimer Disease (AD). Our framework consists of a general approach to extract features from abnormal structures by MRI scans and Electronic Medical Records (EMR), followed by a gradient-based visualization that highlights the positions directly on the original spatial structures. We show that this deep learning framework facilitates the detection of the early preclinical stage of AD that can be used as a reference to assist clinicians.

Introduction Early diagnosis of AD remains a daunting task despite active research in the past two decades. The prevailing diagnostic method mainly relies on the subjects’ behavior and their cognitive development history. Diagnoses can only be made after the subjects demonstrate severe abnormal behaviors, usually after a long period of time. This raises critical concerns/challenges since many patients may have already missed the opportunity for effective early intervention. In this work, we propose a data-driven deep learning framework, named ADNet, to scrutinize and discern abnormal structures from MRI scans of AD patients. After the detection of the structural abnormality and differences, our framework will highlight them over the original images. The overlay process provides guidance and assists clinicians in the early diagnosis of AD. Besides the image data, we also introduce the valuable EMR data as a prior information for more precise diagnosis. We establish a novel approach to build up the graph from the patient information and embed the graph into the framework to predict whether classifying the sample as a potential AD patient.

Methods and Results In the early stage, the abnormal features are not obvious enough to be directly identified by human eyes, while the structural differences are masked behind thousands of voxels of the original images. To unveil the abnormal structural differences, we develop an end-to-end structure called ADNet, to identify the spatial differences between AD and normal control (NC) groups. The framework combines the EMR data and image data for precise predictions. Specifically, based on the patient phenotypes including age, sex, and other cofounders that are highly related with AD, we design a novel graph construction approach to build a patient graph using shared nearest neighbors in canonical correlation vectors that project patient’s information into a correlated low-dimensional space. We then embed the patient information as an input for the prediction. We further investigate the abnormal structure in the MRI data and use the CNN model to learn representative features of brain structure and assign the original images into one of the two groups (AD or NC). A gradient-based visualization method will then highlight the significant features in the original images. To evaluate the selected features, we apply the atlas template AALv3, and compare the features with pre-defined brain areas. Finally, brain areas with highlight features will be compared with previous findings, from which an experienced clinician will check whether the selected areas could be used as biomarkers for AD. To avoid bias, we apply a double-blind test that the clinician is only given the images and patient phenotype without ground truth.

Conclusion We demonstrate ADNet provides an integrated environment for assisting early-stage detection of AD using ADNI dataset. We propose an automotive manner in finding the potential biomarkers that can help the clinicians to diagnose AD in an earlier stage.
Patient Perceptions of Receiving COVID-19 Test Results via an Online Patient Portal

Robert W. Turer, MD MSE,1 Catherine M. DesRoches, DrPH2, Liz Salmi3, Tara Helmer, MPH1, S. Trent Rosenbloom, MD, MPH1

1Vanderbilt University Medical Center, Nashville, TN, USA; 2Beth Israel Deaconess Medical Center, Boston, MA, USA

Introduction

Mounting evidence shows that access to clinical notes empowers patients to better manage their health without burdening clinicians1,2. However, evidence surrounding the immediate release of sensitive test results is controversial3. Vanderbilt University Medical Center (VUMC) has experienced a trend toward increased use of test review features in My Health at Vanderbilt (MHAV), our patient portal4. Historically, VUMC released most results to MHAV within days, only delaying longer for tests with perceived potential for harm if disclosed without the context provided by an interpreting provider. Examples include human immunodeficiency virus (HIV) testing, tumor markers, and genetic testing for Huntington disease.

Patients with COVID-19 have faced stigma in the United States and internationally5, and the disease can cause substantial morbidity and mortality6. As such, we traditionally would have delayed the release of COVID-19 test results to MHAV. However, the overwhelming volume of COVID-19 testing necessitated immediate release of COVID-19 test results to patients after completion7. This online survey-based study evaluated patient perceptions of receiving these results immediately upon completion via MHAV.

Methods

This is a single site online survey-based evaluation of patient perceptions surrounding the release of COVID-19 test results immediately upon completion via MHAV at VUMC - a large, private, nonprofit academic medical center in Nashville, TN. MHAV has more than 730,000 active users representing 62% of clinical encounters since April 1, 2020. This study was deemed exempt by VUMC’s Institutional Review Board. Patients agreeing to direct contact about COVID-19 research opportunities were included in a registry maintained by the Vanderbilt Institute for Clinical and Translational Research (VICTR). From this cohort, we contacted all adult patients with a valid email address who had accessed MHAV since having a resulted COVID-19 test at VUMC.

We partnered with members of the OpenNotes collaborative to design an online survey using REDCap8 to evaluate the following domains: MHAV user role (i.e. patient vs. care partner); ease of use; impact of immediate release; notification of results; impact of viewing results on health management; importance of result sharing; and demographic information.

Results

At VUMC, 25,359 patients who underwent COVID-19 polymerase chain reaction (PCR) testing had logged into MHAV since their first test was ordered. Of this group, 1,390 patients overlapped with the group of patients who had agreed to direct contact about COVID-19 research opportunities and were invited via email to participate. There were 324 (23.3%) responses to the survey. Numbers (percentages) are represented as a fraction of respondents to each question.

MHAV User Role: 310 (96.3%) of respondents were patients themselves, and 290 (90.1%) respondents viewed the results of their COVID-19 test using MHAV.

Ease of use: 262 (95.2%) found the experience “Not very confusing” or “Not confusing at all.” Of the 13 (4.7%) respondents who expressed confusion, 11 reported difficulty interpreting their results and 2 could not find their results.

Impact of Immediate Release: 243 (88.4%) viewed their result before discussing it with a clinician. 99 (40.7%) stated viewing before discussing with a clinician did not change whether they were worried about their health, while 83 (34.2%) felt less worried and 12 (4.9%) felt more worried. 49 (20.2%) were never worried.

Notification of Results: 234 (85.2%) found the default email reminder about their test result “A little useful,” “Somewhat useful” or “Very useful,” while 10 (3.6%) found it “Not at all useful.” 31 (11.3%) respondents reported not seeing an email notification. 119 (43.3%) did not prefer an alternative notification method, while 120 (43.6%) preferred text messages, 58 (21.1%) preferred a phone call from a healthcare professional, and 4 (1.5%) preferred a letter in the mail.

Impact of Viewing Results on Health Management (visual analog scale from 0-100, with 100 being very important):
“...for taking care of your health” - median 86 (IQR 68 – 98).
“...for making next steps for your healthcare” – median 89 (IQR 68 – 99)
“...for helping you feel in control of your care” – median 85 (IQR 67 – 98)

Importance of result sharing: 259 (96.6%) respondents shared their result with someone else - 235 (90.7%) with a family member or relative, 147 (56.8%) with a friend, 80 (30.9%) with a healthcare provider, 24 (9.3%) with a large group of friends (e.g. through social media). Respondents provided additional details mostly related to informing employers or potentially exposed contacts. The importance of COVID-19 result sharing was measured via visual analog scale from 0-100 (100 being very important). The median response was 77 (IQR 50-97). 176 (58.7%) respondents encouraged someone else to get tested because of viewing their result.

Demographics: 284 (95.9%) reported their overall health as good, very good, or excellent. Of 296 respondents, 272 (92.5%) had at least some college education. 16 respondents (5.5%) reported Spanish/Hispanic/Latino ethnicity. 264 (90.7%) respondents identified as white, 14 (4.8%) as black or African American, 1 (0.3%) as American Indian or Pacific Native, 5 (1.7%) as Asian, and 10 (3.4%) as Other. 290 (99.0%) reported speaking English at home. 15 (5.1%) of respondents were unemployed or unable to work. 38 (12.9%) respondents were health care professionals.

Discussion

Our survey indicates that, in our cohort, most patients comfortably viewed and interpreted their COVID-19 results on MHAV prior to any discussion with a clinician without distress, despite concerns about stigmatization or fear of not discussing their result with a clinician. Most were comfortable with email notifications, though some requested text messages or healthcare provider communication. Most respondents reported that viewing their COVID-19 result on MHAV enabled them to take care of their health, make next steps for their healthcare, and feel in control of their care. Sharing of results was important to an overwhelming majority of patients – allowing them to share with a wide range of parties and to encourage someone else to get tested as a result. Other sites may consider providing immediate access to COVID-19 results via online patient portals.

Our limitations are largely demographic in nature: the surveyed cohort was predominantly white, English-speaking, employed, and highly educated. Because we targeted those who had previously responded through electronic means, we may have selected patients with technological literacy compared to the population at large. Further study should evaluate how patient portal-based results sharing impacts patients who are from underrepresented minority groups, who are non-English speaking, or who lack access to technology or higher education.

References
The added value of natural language processing in automated analysis of patient experiences

Marieke M. van Buchem, MSc1, Olaf M. Neve, MD1, Hileen Boosman, PhD1, Ilse M. J. Kant, PhD1, Erik F. Hensen, MD PhD1

1Leiden University Medical Center, Leiden, the Netherlands

Abstract

Patients’ experiences are important to improve healthcare and are often collected using structured questionnaires. Open-ended questionnaires give patients the opportunity to share their unique perspective. Natural language processing (NLP) can automate the analysis of open-ended questionnaires for an efficient and patient-centered approach to improve healthcare. We created a new, open-ended questionnaire, the AI-PREM, and constructed a NLP tool for automatic analysis. We show that this combination leads to broader insights into the patient’s experience.

Introduction

Value-based Health Care (VBHC) aims to increase patient value by improving outcomes and lowering costs. Patient outcomes are a combination of health outcomes and patient experiences. The latter is often measured using structured questionnaires, called patient-reported experience measures (PREMs). Although structured questionnaires are easy to analyze and can lead to valuable insights, they narrow the scope of patient experiences covered. Many PREMs therefore include open-ended questions, giving patients the opportunity to share their unique perspective in more depth. Previous studies have shown that these open-ended questions can complement the structured PREMs, but manual analysis is time-intensive1. Automated analysis would overcome this problem, however patient experiences often contain multiple sentiments and topics, complicating current analyses1,2. Natural language processing (NLP) could be a good technique to address both the identification of topics and the associated sentiments within patient responses to open-ended questions. We created a new open-ended PREM and built a NLP tool to analyze patients’ responses, aiming for a more efficient way to capture a broad spectrum of patient experiences while creating actionable insights.

Methods

The open-ended PREM, called the AI-PREM, contains five open-ended questions about how the patient experienced certain aspects of care delivery. These questions, based on the Picker principles of person centered care3, cover the following aspects: the received information, the organization of care, the collaboration with other healthcare professionals, and how the patient was approached. The phrasing of each of these questions is the same, asking patients how they experienced the specific aspect of care delivery. The last question asks the patient for any other experiences he or she would like to share.

The AI-PREM was sent to patients who visited the vestibular schwannoma integrated practice unit (IPU) at the Leiden University Medical Center in the past 15 years. These patients also filled out a general, structured PREM for comparison.

We developed the NLP tool with a multidisciplinary team consisting of clinicians and technicians. We used an iterative process with short development cycles that always started with a brainstorm session about ways to improve the applicability of the model and ended with feedback on the new output of the tool. The NLP tool that was created during this process consists of three models: two sentiment analysis models using a pretrained language model (Bidirectional Encoder Representations from Transformers, BERT); and a topic model using Non-Negative Matrix Factorization (NMF). See Figure 1 for an overview. For the sentiment analysis models, we manually labeled 60% of the patients’ responses as negative, neutral, or positive per question. A pretrained, multilingual BERT model was finetuned for two binary classification tasks: first classifying texts as negative or non-negative and then classifying the non-negative texts as positive or neutral. These two models were trained on 80% of the labeled data and validated on 20% of the data.

After the sentiment analysis, the data were preprocessed. Spelling was corrected, words were lemmatized, and stop words and words with less than 3 letters were removed. To make the topics as informative as possible we only kept verbs, adverbs, nouns, and adjectives and split the text into n-grams ranging from one to three words. A separate topic model was created per question and sentiment (negative or positive) to improve the interpretability of the extracted topics. This led to 10 separate categories, for which a separate topic model was created using NMF. The number of topics per category was automatically determined using a coherence metric.
To study whether the extracted topics were relevant, the answers of the AI-PREM were compared to the answers of the structured PREM. Furthermore, we plan to ask end-users to determine whether the extracted topics cover the subjects of a random sample of patient responses.

**Preliminary results**

535 patients filled out the AI-PREM questionnaire. The sentiment models classified the texts as negative or non-negative and neutral or positive with an accuracy of 92%. The number of topics per category ranges between two and six and 96.3% of patient responses can be assigned to a topic. We are currently finalizing the comparison to the structured PREM and the manual analysis of the random samples and intend to present this at AMIA. Our preliminary results show that the majority of subjects addressed by patients is covered by at least one of the topics and that the answers to the structured PREM align with the answers to the AI-PREM. Furthermore, the topics provide several new subjects that are not covered in the structured PREM but negatively or positively affect patients’ experience. We also intend to present the current visualization of the output and the collected feedback from end-users during the conference.

**Figure 1.** Overview of the input, models, and output of the AI-PREM tool.

**Discussion**

In this study, we show that an open-ended PREM specifically designed for NLP analysis is able to capture the most important results from a structured PREM, while also capturing various additional topics that are negatively or positively viewed by patients, in a fraction of the time of a manual analysis. This allows patients to elaborate on experiences they deem important, yielding information additional to structured PREMs and more actionable insights specific for the IPU. Furthermore, the use of an unsupervised approach makes the method easily applicable and potentially generalizable to other settings. Future research and implications of this study will be discussed at AMIA.

**References**

The GA4GH Variation Representation Specification (VRS): a Computational Framework for the Precise Representation of Molecular Variation

Alex H. Wagner, PhD1,2, Lawrence Babb, BSEE3, Melissa Cline, PhD4, Helen Schuilenburg, BSc5, Tristan Nelson, BA6, Robert R. Freimuth, PhD7, Reece K. Hart, PhD8

1Nationwide Children's Hospital, Columbus, OH, USA; 2The Ohio State University, Columbus, OH, USA; 3Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA; 4UC Santa Cruz Genomics Institute, Santa Cruz, CA, USA; 5European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, United Kingdom; 6Geisinger Health, Danville, PA, USA; 7Center for Individualized Medicine, Mayo Clinic, Rochester, MN, USA; 8MyOme, Inc., San Carlos, CA, US

Introduction
Ensuring that precision genomic medicine is effective for individuals and for health systems requires clinicians, researchers, and testing laboratories to communicate genomic variation and related information reliably. Although widely-adopted standards for certain classes of variation already exist, many of these formats have been purpose-built for specific applications, including human-readable standards such as the Human Genome Variation Society (HGVS) variant nomenclature, the International System of Human Cytogenomic Nomenclature (ISCN), and the PharmVar Pharmacogenetics nomenclature, as well as genome-oriented flat file formats such as the Variant Call Format (VCF), among others. All current standards have design constraints that preclude a comprehensive coverage of variation types and extensibility to new types, including highly complex forms of variation.

Methods
In response to this need, the Global Alliance for Genomics and Health (GA4GH) Genomic Knowledge Standards (GKS) Work Stream led the development of the Variation Representation Specification (VRS, pronounced “verse”), a community-driven and extensible specification to standardize the exchange of diverse and complex variation data. In the course of development, the standard was applied and tested in real-world data exchange systems hosted by GA4GH Driver Projects, including the Clinical Genome (ClinGen) resource, the Variant Interpretation for Cancer Consortium (VICC), and the BRCA Exchange.

Results
VRS is a computational framework encompassing interdependent components for the precise representation of variation: a formal terminology, an extensible information model, schema, and conventions for normalizing and identifying variation. It is more verbose than other contemporary human-readable variant representations, but better suited to expressing and representing complex variation concepts for computational use. It is also a natural complement of human-readable nomenclatures when used for the exchange of genomic information from databases, clinical reports, or scientific manuscripts. VRS currently covers many classes of variation that are defined in cis on a contiguous molecule such as single nucleotide variants (SNVs), multi-nucleotide variants (MNVs), indels, repeats, and haplotypes.

VRS also provides semantic clarity to more complex forms of variation. For example, due to design constraints in other variant representation systems, resources often represent these variants as a conflation of the similar but distinct concepts of systemic copy variation and tandem repeating elements within a molecule. In addressing this, systemic copy number and repeated sequence expressions were created to precisely express these distinct concepts. This complements existing cross-community concepts in VRS, including variation describing giemsa-stained chromosome banding patterns. Additional progress has been made in the representation of gene fusion and structural variation events, a major objective of the VRS version 1.3 release.

A key design decision of VRS is the representation of Variation as Value Objects: minimal information objects intrinsically defined by their content (e.g. a C>T transition in residue 2 of the sequence CCTAC), rather than records for human presentation from system identifiers (e.g. clinvar:13961). This design choice is fundamental to each component of the specification; objects are immutable and have a specific meaning prescribed by the attributes they contain. Values extraneous to the meaning of a VRS object, such as the name of a location (“chr 7”) or a label for a
well-known variant (“BRAF V600E”) are not captured directly within VRS objects. Combined with a process for normalizing variation objects by implementing the NCBI Variant Overprecision Correction Algorithm (VOCA)(1), this provides a precise, singular representation of a given variation concept. Normalized value object representation also enables the use of the associated GA4GH Computed Identifier Algorithm, a novel mechanism for the federated identification of molecular variation data (Figure 1).

**Figure 1.** VRS provides a mechanism for federated variation identification via the Computed Identifier Algorithm. (A) The Computed Identifier Algorithm is defined in three stages. First, an identifiable VRS object such as an Allele (blue box) is transformed into a well-defined and canonical serialized JSON representation. The serialized Binary Large Object (BLOB) is then digested via the SHA-512 algorithm, truncated to retain only the first 24 bytes, and subsequently encoded using base64url. The resulting digest string (green text) is then appended to the object type identifier; for an Allele object, the identifier prefix is “VA” (blue text). The identifier is then assembled into a compact URI (CURIE) under the ga4gh namespace (orange text). (B) Use of the VRS framework enables de-duplication of identical variation concepts with differing HGVS descriptions. Here, multiple synonymous HGVS descriptions are indicated for a variant on genome builds GRCh37 and GRCh38, the corresponding transcript variant, and predicted protein translation. These four contexts (two genome assemblies, transcript, and protein) resolve to four distinct identifiers, regardless of which synonymous description is used to build the VRS object. Ellipses (“...”) used in objects and strings in this diagram represent content that is omitted for simplicity of presentation. For additional details, see [https://vrs.ga4gh.org/en/stable/impl-guide/computed_identifiers.html](https://vrs.ga4gh.org/en/stable/impl-guide/computed_identifiers.html).

Application of VRS to Driver Project systems revealed novel challenges created by the value object design decision for VRS Variation. Specifically, transmission of human-readable identifiers for genomic variation is a common use case in contemporary molecular variation databases, underlining the need to have complementary formats and nomenclatures for variation description. In addition, we identified that while VRS provides a novel solution to the challenge of federated variation identification, it is beyond the remit of VRS to specify methods or policies for simplifying aggregate variation (e.g. a ClinVar Variation record aggregating multiple genomic, transcript, and protein contexts, such as those illustrated in Figure 1B). To address these concerns we developed VRS Added Types for Interoperable Loquacious Exchange (VRSATILE; [vrsatile.readthedocs.io](http://vrsatile.readthedocs.io)), a set of proposed extensions and a community guide to maximizing the utility of VRS in real-world data exchange.

**Conclusion**

VRS is a framework for the computational representation of molecular variation. The extensibility of this framework enables the representation of disparate and complex classes of variation, including systemic copy number variation, cytogenetic variation, gene product expression, gene fusions, and structural variation. VRS has demonstrated value in production genomic data exchange systems, and is complemented by proposed extensions in VRSATILE.

**References**

An Evaluation of Racial Disparity in 30-Day Hospital Readmission Models

H. Echo Wang, MSPH1, Hadi Kharrazi, MD, PhD1
1Johns Hopkins University School of Public Health, Baltimore, MD

Introduction
The potential for predictive algorithms to exacerbate inequity in healthcare has been increasingly recognized.1 Efforts to quantitatively measure bias have also surged in response to the rising concern of algorithmic inequity.2 However, bias measurement methods are rarely reported due to their poor interpretability and mathematical incompatibilities between notions of bias.3 To address the lack of bias measurement, this study aims to explore the potential bias of common 30-day hospital readmission models in generating disparate results for African American vs. White, and for low-income vs. high-income patients.

Methods
Study Population: The retrospective study included 1.5 million adult inpatient discharges 2016-2018 in Maryland. Study Outcome: Unplanned 30-day hospital readmissions were identified using the methodology adopted by CMS.3 The unit of analysis was hospital discharge. If more than one readmission occurred within 30 days, only one readmission was counted. Predictive Models: LACE index,4 HOSPITAL score5, and CMS hospital-wide all-cause readmission measure6 were included in the analysis. The CMS measure was applied “as is” with existing coefficients, or “retrained” using 50% of the sample. Measures: We measured bias between black and white races, and between low-income and high-income subpopulations. Predictive performances of each model were derived for all population and each subpopulation using AUC, Brier statistic, and Hosmer-Lemeshow goodness of fit. Common mathematical bias measures were calculated for each model: statistical parity (equal total positive rates), equal opportunity (equal true positive rates), false positive rate parity (equal false positive rates), and false negative rate parity (equal false negative rates). Generalized entropy index (GEI), which was proposed to measure algorithmic fairness, was also calculated.7 To calculate these measures, standard model thresholds were used: 10 points for LACE, high-risk (5 in the adjusted scoring) for HOSPITAL, and the ideal threshold derived using the receiver operating characteristic (ROC) curve for the CMS measures. Statistical Analysis: Primary analyses were conducted in R. Aggregate condition categories were calculated in SAS. The results were stratified by hospitals serving mostly blacks vs. whites. Sensitivity analysis was performed to explore how different thresholds affect the model accuracy and bias.

Results
Demographics: Among the 829,584 unique patients, blacks and low-income groups were more likely to be on public insurance, have longer length of inpatient stay, more inpatient procedures, higher inpatient charges, higher emergency admissions and more comorbidities. Blacks were more likely to live in urban areas and hospitalized in major teaching or large hospitals (>=200 beds). Low-income groups were more likely to live in rural areas and hospitalized in major teaching or large hospitals. Observed and Estimated Rates: The observed 30-day unplanned readmission rates were 11.4% for all population, 11.2% for white, 13.0% for black, 10.6% for high-income and 12.9% for low-income populations. LACE in general overestimated readmission rates but more so in blacks and low-income groups. The readmission rates estimated by HOSPITAL score are close to observed rates. The CMS as-is model underestimated across subpopulations and the estimated rates were relatively flat, while the retrained CMS model overestimates in all subpopulations to a similar degree. Predictive Performance: For each model, predictive performance measures remained largely consistent across subpopulations. LACE, HOSPITAL and CMS as-is models had moderate predictive performance (AUC 0.62-0.69), but retrained CMS model improved the performance across subpopulations (AUC 0.73-0.74). Bias Measures: Misclassification rates (i.e. FPR parity, and FNR parity) indicated relative between-group bias between models, while the absolute between-group difference (e.g. FPR difference between black and white) may not directly translate to bias as the observed readmission rates vary between groups (Table 1). Retrained CMS model and HOSPITAL score had the lowest racial bias based on FPR parity and FNR parity. The bias between income groups was low across models and lowest in the retrained CMS model. Based on GEI, Retrained CMS model and HOSPITAL score have the lowest GEI indicating the least bias, while CMS as-is model has the highest bias. When stratified by majority black and majority white hospital, the racial bias diminishes in majority black hospitals, in which the observed readmission rates are close between black and white. Sensitivity Analysis: Several trade-offs were observed. The commonly used thresholds of LACE and HOSPITAL maximized accuracy but were not fair between subpopulations. A lower threshold improved fairness but sacrificed accuracy. The thresholds of the two CMS
models were selected based on ROC curve thus minimized overall FNR and FPR; however, a clear threshold to simultaneously improve both bias measures were lacking.

### Table 1: False positive rate (FPR) parity, false negative rate (FNR) parity and generalized entropy index (GEI)

<table>
<thead>
<tr>
<th>Model</th>
<th>White</th>
<th>Black</th>
<th>Ratio (B/W)</th>
<th>Diff (B-W)</th>
<th>High-income</th>
<th>Low-income</th>
<th>Ratio (H/L)</th>
<th>Diff (H-L)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LACE</td>
<td>FNR</td>
<td>0.70</td>
<td>0.59</td>
<td>0.85</td>
<td>0.69</td>
<td>0.62</td>
<td>0.90</td>
<td>-0.07</td>
</tr>
<tr>
<td></td>
<td>FPR</td>
<td>0.12</td>
<td>0.18</td>
<td>1.43</td>
<td>0.05</td>
<td>0.12</td>
<td>1.73</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>GEI</td>
<td>--</td>
<td>--</td>
<td>0.030</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>0.019</td>
</tr>
<tr>
<td>HOSPITAL</td>
<td>FNR</td>
<td>0.76</td>
<td>0.70</td>
<td>0.92</td>
<td>0.76</td>
<td>0.70</td>
<td>0.92</td>
<td>-0.06</td>
</tr>
<tr>
<td></td>
<td>FPR</td>
<td>0.10</td>
<td>0.11</td>
<td>1.17</td>
<td>0.02</td>
<td>0.09</td>
<td>1.29</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>GEI</td>
<td>--</td>
<td>--</td>
<td>0.005</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>0.007</td>
</tr>
<tr>
<td>CMS (as is)</td>
<td>FNR</td>
<td>0.34</td>
<td>0.24</td>
<td>0.70</td>
<td>0.32</td>
<td>0.26</td>
<td>0.81</td>
<td>-0.06</td>
</tr>
<tr>
<td></td>
<td>FPR</td>
<td>0.42</td>
<td>0.53</td>
<td>1.27</td>
<td>0.11</td>
<td>0.44</td>
<td>1.15</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td>GEI</td>
<td>--</td>
<td>--</td>
<td>0.048</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>0.026</td>
</tr>
<tr>
<td>CMS(retrain)</td>
<td>FNR</td>
<td>0.31</td>
<td>0.26</td>
<td>0.84</td>
<td>0.30</td>
<td>0.27</td>
<td>0.88</td>
<td>-0.04</td>
</tr>
<tr>
<td></td>
<td>FPR</td>
<td>0.30</td>
<td>0.35</td>
<td>1.15</td>
<td>0.04</td>
<td>0.30</td>
<td>1.12</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>GEI</td>
<td>--</td>
<td>--</td>
<td>0.019</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>0.014</td>
</tr>
</tbody>
</table>

### Discussion

The bias measures related to misclassification were helpful to compare bias between models. Retraining of a model is effective in reducing bias between groups. In addition, as confusion-matrix based metrics are driven by baseline prevalence rates, some proposed bias measures (e.g. equal positive rates) are determined to be inappropriate for the predictive task due to the significant difference in observed readmission rates between groups. For the same reason, the absolute ratio or difference of the misclassification measures should not be simply compared to 1 (or 0) to draw a conclusion regarding bias. When the observed readmission rates were similar between groups, as in the case of majority black hospitals, the magnitude of bias decreased or disappeared, despite that patient-level predicted risk remained the same. Further analysis on misclassified cases showed blacks were sicker than white at the same predicted score; false negative (predicted low risk but readmitted) blacks were sicker when they readmit; and false positive (predicted high risk but not readmitted) blacks were more likely to be hospitalized later (black: 65%, white 59%). It is apparent that blacks need more healthcare intervention across risk levels, including those falsely classified as low risk. Therefore, quantitative bias measures were helpful to compare models under the same condition, but have limited utility in determining the direction, magnitude or implications of bias.

The sensitivity analysis also illustrated the issue of selecting eligibility threshold of predictive models. Regardless of the type of predictive models, it is a common practice to determine an eligibility threshold post-modeling. Classic decision theories do not account for group bias in determining the threshold, but our work demonstrated the danger of creating inequity with slight threshold change. We recognize the selection of the threshold with many competing factors (e.g., accuracy, benefit, harm, bias) is not a straightforward task. However, it is increasing important to recognize and monitor the differential effect between groups, and to seek appropriate adjustments in practice.

### References

4. van Walraven C, Dhalla IA, Bell C, et al. Derivation and Validation of An Index to Predict Early Death or Unplanned Readmission After Discharge From Hospital to The Community. *CMAJ* 2010.
Li Qin Wang, PhD1,2, Dinah Foer, MD1,2, Erin MacPhaul, MS1, Ying-Chih Lo, MD1,2, David W. Bates, MD, MSc1,2, Li Zhou, MD, PhD1,2
1Brigham and Women’s Hospital, Boston, MA, USA; 2Harvard Medical School, Boston, MA, USA

Introduction
As of July 2021, there had been over 190 million confirmed cases of the coronavirus disease 2019 (COVID-19) and four million deaths. Emerging prospective and retrospective studies suggest that some patients have persistent symptom, and/or develop delayed or long-term complications after their recovery from acute COVID-19, which may also be referred to as post-acute sequelae of SARS-CoV-2 infection (PASC) syndrome or long COVID.1,2 The onset, scope and duration of these symptoms represent a new phase of the pandemic, with significant implications for health care delivery. With rapid evolution of the scientific literature, characterization of PASC symptoms varied widely by study.1 This heterogeneity of research finding is attributed to multiple factors, including variation in study design, patient populations, “post-acute COVID” timeframe, and sample size.2 Most early studies on PASC symptoms relied on patient survey data, manual chart review, and in person follow-up. These studies were often limited by sample size and reporting biases. Longitudinal electronic health record (EHR) data serve as a rich data source for studying PASC symptoms. However, patient symptoms are predominantly found in unstructured, free text data in clinical notes which require specialized natural language processing (NLP) tools for search and identification.3 NLP to extract long-term symptoms in the post-acute COVID-19 period remains an unmet need. One big challenge is the wide variation in potential PASC symptoms. Therefore, the objective of this study was to develop a comprehensive PASC symptom lexicon by leveraging real-world clinical EHR data, medical ontologies, an NLP tool and manual chart review.

Methods
The study was conducted in the Mass General Brigham (MGB) healthcare system, the largest integrated healthcare delivery system in New England. We used a two-phase approach to develop and evaluate a post-acute COVID-19 symptom lexicon (Figure 1). In the first phase of this study, we focused on lexicon development. First, we identified patients who were ≥18 years of age and had a positive test result for SARS-CoV-2 by polymerase chain reaction (PCR) clinical assay between March 4, 2020 and February 09, 2021. We defined the post-acute COVID-19 period as the days on and after day 51 from first positive PCR test and used all clinical notes from days 51-100 for each patient. Due to the focus on post-acute COVID-19 symptoms, patients who died before day 51 were excluded from the study. We divided the corpus of clinical notes into development (90% of the study cohort) and validation (10% of the study cohort) datasets to meet our study objectives. Second, using the Unified Medical Language System® (UMLS) and SNOMED CT, we compiled an initial symptom lexicon containing selected UMLS concepts and related synonyms. Third, we applied the MTERMS NLP tool4 to extract mentions of the symptoms in the lexicon from the development dataset. Fourth, we refined the lexicon based on an iterative process including: (1) symptom ranking based on prevalence; (2) manual review to identify post-acute COVID-19 symptoms; (3) consolidation of concepts of similar meaning; and (4) evaluation of NLP performance. Once the lexicon reached a satisfactory level of performance during the NLP evaluation in the development dataset, we moved on to the second phase, which evaluated its performance in the NLP system using the validation dataset of a different cohort to gain insights about the lexicon. We measured performance in terms of precision (or positive predictive value) for the 50 most common symptoms and recall (or sensitivity) across all the symptoms. The final symptom lexicon and NLP were used to evaluate NLP performance in the validation dataset. After validation, we applied the lexicon and NLP to the entire clinical note corpus. We calculated the frequency of PASC symptoms within the entire study population during the 2-month follow-up.

Figure 1. Schematic diagram of post-acute COVID-19 symptom development and natural language processing (NLP) system evaluation.
Results

The development dataset contained 299,140 notes from 23,505 patients (14,578 [62.0%] female; mean [SD] age 51.6 [18.2] years). The validation dataset contained 29,739 notes from 2,612 patients (1,599 [61.2%] female; mean [SD] age 51.4 [18.4] years). The initial lexicon included a total of 157,245 unique concepts and 604,056 synonyms. Eleven percent (n=17,701) of the concepts were mentioned in the development dataset. After symptom review and consolidation, the final post-acute COVID-19 symptom lexicon included 355 post-acute COVID-19 symptoms consolidated from a total of 1,520 UMLS concepts. Table 1 shows the most common 50 symptoms in the entire patient cohort and the NLP performance in clinical note symptom extraction for individual symptoms measured in the validation dataset. 46 concepts (92%) had precision measured above 0.90; average precision was 0.94 (range, 0.82 to 1.0). For recall, a total of 1,481 sentences were reviewed from 50 randomly selected notes. Manual review identified 104 symptom terms, among which NLP identified 87 symptoms; therefore, estimated recall was 0.84.

Discussion and Conclusion

This study detailed the development of a comprehensive PASC symptom lexicon from the EHR, and validation of an NLP tool using the lexicon for symptom extraction. Free-text clinical notes represent an underutilized data source in the active field of post-acute COVID-19 research. This work advances the study of PASC symptoms by providing a systematic approach and scalable tool to identify patient PASC symptoms in large datasets. These symptoms can be used to characterize patient populations or organ/system-based domains for targeted preventative or therapeutic interventions, and support future research.

References


<p>| Table 1 | 50 most common post-acute COVID-19 patient symptoms in EHR clinical notes by symptom frequency, and corresponding precision of natural language processing performance for unique symptom extraction |</p>
<table>
<thead>
<tr>
<th>Top 1-25 Symptoms</th>
<th>Prevalence (%)</th>
<th>Precision</th>
<th>Top 26-50 Symptoms</th>
<th>Prevalence (%)</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pain</td>
<td>43.1</td>
<td>0.94</td>
<td>Insomnia</td>
<td>11.2</td>
<td>0.94</td>
</tr>
<tr>
<td>Anxiety</td>
<td>25.8</td>
<td>0.98</td>
<td>Pain in extremities</td>
<td>10.7</td>
<td>1.0</td>
</tr>
<tr>
<td>Depression</td>
<td>24.0</td>
<td>0.90</td>
<td>Paresthesia</td>
<td>10.7</td>
<td>0.92</td>
</tr>
<tr>
<td>Fatigue</td>
<td>23.4</td>
<td>1.0</td>
<td>Peripheral edema</td>
<td>10.5</td>
<td>0.98</td>
</tr>
<tr>
<td>Joint pain</td>
<td>21.0</td>
<td>0.98</td>
<td>Palpitations</td>
<td>10.3</td>
<td>0.94</td>
</tr>
<tr>
<td>Shortness of breath</td>
<td>20.8</td>
<td>0.94</td>
<td>Diarrhea</td>
<td>10.3</td>
<td>0.92</td>
</tr>
<tr>
<td>Headache</td>
<td>20.0</td>
<td>0.92</td>
<td>Itching</td>
<td>9.4</td>
<td>0.92</td>
</tr>
<tr>
<td>Nausea and/or vomiting</td>
<td>19.9</td>
<td>1.0</td>
<td>Erythema</td>
<td>9.2</td>
<td>0.98</td>
</tr>
<tr>
<td>Myalgia</td>
<td>19.0</td>
<td>0.96</td>
<td>Lower urinary tract symptoms</td>
<td>8.7</td>
<td>0.98</td>
</tr>
<tr>
<td>Gastroesophageal reflux</td>
<td>18.6</td>
<td>0.94</td>
<td>Lymphadenopathy</td>
<td>8.3</td>
<td>0.96</td>
</tr>
<tr>
<td>Cough</td>
<td>17.5</td>
<td>0.92</td>
<td>Edema</td>
<td>7.9</td>
<td>0.88</td>
</tr>
<tr>
<td>Back pain</td>
<td>16.9</td>
<td>0.98</td>
<td>Weight gain</td>
<td>7.3</td>
<td>0.98</td>
</tr>
<tr>
<td>Stress</td>
<td>15.1</td>
<td>0.86</td>
<td>Sinonasal congestion</td>
<td>7.1</td>
<td>0.96</td>
</tr>
<tr>
<td>Fever</td>
<td>14.7</td>
<td>0.94</td>
<td>Pain in throat</td>
<td>6.4</td>
<td>0.98</td>
</tr>
<tr>
<td>Swelling</td>
<td>14.7</td>
<td>0.90</td>
<td>Abnormal gait</td>
<td>5.9</td>
<td>1.0</td>
</tr>
<tr>
<td>Bleeding</td>
<td>14.7</td>
<td>0.90</td>
<td>Respiratory distress</td>
<td>5.8</td>
<td>0.82</td>
</tr>
<tr>
<td>Weight loss*</td>
<td>14.2</td>
<td>0.98</td>
<td>Visual changes</td>
<td>5.8</td>
<td>0.92</td>
</tr>
<tr>
<td>Abdominal pain</td>
<td>14.1</td>
<td>0.98</td>
<td>Chills</td>
<td>5.6</td>
<td>0.86</td>
</tr>
<tr>
<td>Dizziness or vertigo</td>
<td>14.0</td>
<td>0.94</td>
<td>Urinary incontinence</td>
<td>5.6</td>
<td>0.96</td>
</tr>
<tr>
<td>Chest pain</td>
<td>12.5</td>
<td>0.90</td>
<td>Sleep apnea</td>
<td>5.4</td>
<td>0.94</td>
</tr>
<tr>
<td>Weakness</td>
<td>12.3</td>
<td>0.94</td>
<td>Confusion</td>
<td>5.4</td>
<td>0.98</td>
</tr>
<tr>
<td>Constipation</td>
<td>11.9</td>
<td>0.96</td>
<td>Hearing loss</td>
<td>5.2</td>
<td>1.0</td>
</tr>
<tr>
<td>Skin lesion</td>
<td>11.9</td>
<td>0.94</td>
<td>Problem with smell or taste</td>
<td>5.0</td>
<td>0.94</td>
</tr>
<tr>
<td>Wheezing</td>
<td>11.9</td>
<td>0.98</td>
<td>Dysphagia</td>
<td>4.9</td>
<td>0.98</td>
</tr>
<tr>
<td>Rash</td>
<td>11.4</td>
<td>0.82</td>
<td>Loss of appetite</td>
<td>4.8</td>
<td>0.96</td>
</tr>
</tbody>
</table>
Comparison of Machine Learning Algorithms for Earlier Detection of Cognitive Decline from Clinical Notes in the Electronic Health Records

Li Qin Wang, PhD1,2, John Laurentiev, MS1, Jie Yang, PhD1,2,3, Ying-Chih Lo, MD1,2, Rebecca E. Amariglio, PhD1,2,4, Gad A. Marshall, MD1,2,4, Li Zhou, MD, PhD1,2
1Brigham and Women’s Hospital, Boston, MA, USA; 2Harvard Medical School, Boston, MA, USA; 3Zhejiang University, Hangzhou, Zhejiang, China; 4Massachusetts General Hospital, Boston, MA, USA

Introduction

There are nearly 6 million people diagnosed with Alzheimer’s disease (AD) at the stage of dementia in the US, and the prevalence rises dramatically with age.1 Mild cognitive impairment (MCI) and subjective cognitive decline (SCD) represent precursor stages that can serve as targets for early treatment.2 Early detection of cognitive decline can facilitate participant enrollment in clinical trials and early interventions. However, challenges exist in the early detection of patients with cognitive decline. There are insufficient numbers of specialists with expertise in diagnosing cognitive decline to see all at-risk patients. Instead, primary care physicians and other non-dementia specialists have direct contact with these patients but not necessarily the time or tools needed to diagnose them. Systematically reviewing large electronic health records (EHR) data collected across patients’ full visit history within the healthcare system can facilitate early detection of cognitive decline by tracking back to when patients reported their first signs or symptoms of cognitive decline to any healthcare provider. This in turn may help trigger a more detailed evaluation in primary care settings and beyond, facilitate participant enrollment in clinical trials and early interventions. Given that self-reported concerns about cognitive status do not imply a diagnosis of cognitive decline by a healthcare professional, cognitive symptoms may be merely documented in provider notes. Current approaches to identifying cognitive decline rely on billing codes, which can be insensitive; manually curated keywords, which can be limited and lack accuracy; or manual medical record review, which is costly and non-scalable. Therefore, in this study, we aim to develop and compare machine learning models to detect any evidence of cognitive decline from clinical notes.

Methods

Setting, data sources, and study sample. This study was conducted at Mass General Brigham (MGB, formerly Partners Healthcare), a large, integrated healthcare delivery system in Greater Boston, Massachusetts. We used data from MGB’s enterprise data warehouse (EDW). We first identified patients aged 50 and above with an initial diagnosis of MCI (ICD 10: G31.84) between January 01, 2019 and January 01, 2020. We extracted clinical notes documented four years before the initial MCI diagnosis. Due to the fact that clinical notes can be very long and contain many sections that are not relevant to cognitive assessments (such as the allergy section), we segmented clinical notes into sections. That is, our classification task was to identify whether a note section indicates that a patient has cognitive decline. We used the Medical Text Extraction, Reasoning, and Mapping System (MTERMS) natural language processing (NLP) system to split clinical notes into sections.3

Definition of cognitive decline. This classification task aimed to identify patients at any stage of cognitive decline, ranging from SCD to MCI to dementia. Therefore, cognitive decline can be a cognitive concern, symptoms (e.g., memory loss), diagnoses (e.g., MCI, AD dementia), cognitive assessments (e.g., mini-cog), or cognitive-related therapy or treatments (e.g., cognitive-linguistic therapy). We focused on progressive cognitive decline that is likely to be consistent with or lead to MCI, while we excluded cases (assigned negative labels) that were less likely progressive (e.g., cognitive function has improved), or transient (e.g., temporarily forgetful or occasional memory loss due to medication intake), or cognitive function affected soon after some event (e.g., surgery, injury, stroke), as the cognitive decline might be reversible with a return to normal cognition. We also excluded cases which had broader or uncertain indication of cognitive decline (e.g., altered mental status).

Reference datasets. We manually labeled sections of notes for cognitive decline and created two gold standard datasets. To increase the positive case density of the reference dataset, we applied a list of keywords that were initially identified from cognitive-related topics based on topic modeling and enriched with experts knowledge a priori, to filter for sections that likely contain indications of cognitive decline, including memory delirium dementia psych* neuro* mental Alzheimer confus* mood cognit* forget* agitat* MoCA montreal MMSE remember difficult recall function word evaluat* score drive attention mild impairment speech question disorientation orientation sleep alter

1548
exam decline worse loss. Three annotators (i.e., LW, JL, and YL) achieved a good agreement with a Fleiss’ kappa of 0.83 in labeling sections of clinical notes for cognitive decline. Dataset I contained a random sample of 4,950 note sections filtered by the keywords and was used for model development and testing. Dataset II contained 2,000 randomly selected sections without keyword filtering for assessing whether the model performance was dependent on specific keywords. Any cases for which labeling was uncertain were resolved by the domain experts.

**Model development and validation.** We implemented a hierarchical attention-based deep learning structure and four baseline machine learning algorithms, including logistic regression (LR), random forest (RF), support vector machine (SVM), and XGBoost. The deep learning algorithm was developed in a prior study that incorporates a convolutional neural network (CNN), recurrent neural network (RNN), and attention layers for the purpose of handling word variations, context and interpretation of the prediction respectively. In the deep learning model, each section was regarded as a sequence of tokens, with individual words represented by word embeddings, for which we used a word2vec approach and trained 100-dimensional embeddings on a large corpus of 3,729,838 notes from 10,837 patients with an initial MCI diagnosis between January 2017 and January 2020. In the four machine learning baseline models, each section was represented with term frequency-inverse document frequency vectors based on n-grams (where n=1). The algorithms were trained and tested through 5-fold cross validation using dataset I and further assessed in dataset II. We compared the deep learning model with the four baselines in cognitive decline detection in terms of area under the receiver operating characteristic curve (AUROC) and area under the precision-recall curve (AUPRC).

**Table 1. Performance of Five Machine Learning Models for Detecting Cognitive Decline from Clinical Notes**

<table>
<thead>
<tr>
<th>Model</th>
<th>Dataset I (4,950 Sections)</th>
<th>Dataset II (2,000 Sections)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AUROC (95% CI)</td>
<td>AUPRC (95% CI)</td>
</tr>
<tr>
<td>LR</td>
<td>0.936 (0.929-0.943)</td>
<td>0.880 (0.867-0.893)</td>
</tr>
<tr>
<td>RF</td>
<td>0.950 (0.944-0.956)</td>
<td>0.889 (0.875-0.902)</td>
</tr>
<tr>
<td>SVM</td>
<td>0.939 (0.933-0.946)</td>
<td>0.883 (0.869-0.897)</td>
</tr>
<tr>
<td>XGBoost</td>
<td>0.953 (0.946-0.960)</td>
<td>0.882 (0.864-0.900)</td>
</tr>
<tr>
<td>Deep Learning</td>
<td>0.971 (0.967-0.976)</td>
<td>0.933 (0.921-0.944)</td>
</tr>
</tbody>
</table>

models. Compared to the four baseline models (i.e., RF, LG, SVM and XGBoost), the deep learning model achieved the best performance, with an AUROC of 0.971 and AUPRC of 0.933 in dataset I, an improvement of 0.018 in AUROC and 0.051 in AUPRC compared to the best baseline model, i.e., XGBoost. The model is generalizable to the dataset regardless of keywords with an AUROC of 0.997 and an AUPRC of 0.929, an increase of 0.009 in AUROC and 0.031 in AUPRC compared to XGBoost. With a 0.5 probability as the cutoff, the deep learning model could achieve a positive predictive rate (PPV) of 0.848 and a sensitivity of 0.925 in dataset I and a PPV of 0.771 and a sensitivity of 0.928 in dataset II.

**Discussion and Conclusion**

Early detection of cognitive decline offers promise to facilitate patient recruitment from EHR for clinical trials or identify those for closer screening in primary care visits. A machine learning approach can be used to automatically identify evidence of cognitive decline from thousands or millions of clinical notes along patients’ longitudinal EHR earlier than it may be detected in structured EHR fields like diagnosis codes, medications, and problem lists. This study demonstrates that a deep learning model that was trained on a small set of labeled sections is effective and reliable in detecting cognitive decline. In the future, we plan to apply the deep learning model to identify patients’ first evidence of cognitive decline in the EHR, to facilitate antecedent risk factor and outcome analyses.

**Acknowledgement:** This study was supported by Brigham and Women’s Hospital Women’s Brain Initiative.

**References**

The impact of COVID-19 pandemic on health information sharing and patient-generated health data: findings from the Health Information National Trends Survey 2020

Zidan Wang¹, Jiancheng Ye²*

¹ Department of Statistics, Northwestern University, Evanston, IL, USA; ² Feinberg School of Medicine, Northwestern University, Chicago, IL, USA

Introduction
The emerging health technologies and digital services provide effective ways of collecting human behavior information, gathering patient-generated health data (PGHD), and sharing health-related information outside of clinical settings in a systematic way, and thus making interventions timely. PGHD can offer a broader view of the patient experience while generating real-time, continuous streams of health-related data before, during, and after treatment, or as a more general study of disease. Coupled with population health informatics tools, these technologies can track people's digital exhaust, which includes PGHD and social network platform use. The internet and social networking services have become important sources of health-related information on COVID-19 and on critical responses. Virtual health – compounded by the proliferation of sensors, mobile apps, monitors, digital therapeutics, at-home diagnostics, patient surveys, social channels, and other monitoring tools, can be used to make better health care decisions. This study aims to describe and compare characteristics of the population who participated in a national cross-sectional survey before and after the COVID-19 pandemic, including physical health, sleep, and alcohol use. We will also examine the patterns of social networking service use, patient-generated health data on the digital platforms, and health information sharing attitudes and activities.

Methods
Data used in this study were from the fourth round of data collection for HINTS 5 (Cycle 4), which began in February 2020 and concluded in June 2020. Because data collection for Cycle 4 started before COVID-19 became an international pandemic and continued after the pandemic was declared by the World Health Organization, the variable PANDEMIC was created to flag respondents whose survey was received after the COVID-19 pandemic declaration (March 11, 2020). This variable will facilitate the examination of responses before and after the COVID-19 pandemic that became a widespread issue of concern in the United States. The sample was divided into two groups by the dichotomous variable (PANDEMIC). We used participants' self-reported information on age, sex, race, ethnicity, level of education, income, and usual source of care as our sociodemographic variables. We used the survey package in R programming language 5 (R Foundation, Vienna, Austria) to account for the complex sampling design used in HINTS and incorporated the Taylor series (linear approximation) to generate accurate variance estimation. All analyses used weighted data to provide nationally representative estimates. The pairwise deletion was used to deal with missing data to preserve more information. To assess the characteristics of sociodemographic, general health, chronic diseases, social networking use, alcohol consumption, and sleeping variables, we generated weighted 2-way crosstab tables, which were tested with a Pearson Chi-square test of association.

The univariate logistic regression was built to examine the association between each predictor and mental health. Then we used multivariate logistic regression analyses using survey-weighted generalized linear modeling function in R. Odds ratio (OR) and 95% confidence interval (CI) for both models were presented. All reported P-values were two-tailed and the cutoff of P<0.05 was used to determine statistical significance for all analyses.

Results
The final HINTS 5, Cycle 4 (2020) sample consists of 3,865 respondents. Post-pandemic respondents were more likely to be younger (P<0.001), Asian or Black (P=0.001), and Hispanic (P=0.033). They were also less likely to have a usual source of care (P<0.001) or a cancer history (P=0.045). There were no significant differences between the pre-pandemic and post-pandemic groups regarding other sociodemographic and clinical characteristics.
Table 1. Prevalence estimates for characteristics of health information sharing and social networking service use.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Unweighted % (N=3865)</th>
<th>Weighted % (N=253,815,197)</th>
<th>Pre-pandemic, Weighted % (N=89,632,202)</th>
<th>Post-pandemic, Weighted % (N=164,182,995)</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Used electronics to look for health information, %</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>69.9</td>
<td>72.6</td>
<td>76.0</td>
<td>70.8</td>
<td>0.029</td>
</tr>
<tr>
<td>No</td>
<td>30.1</td>
<td>27.4</td>
<td>24.0</td>
<td>29.2</td>
<td></td>
</tr>
<tr>
<td>Used electronics to look up medical test results, %</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.018</td>
</tr>
<tr>
<td>Yes</td>
<td>43.0</td>
<td>42.2</td>
<td>40.0</td>
<td>46.3</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>57.0</td>
<td>57.8</td>
<td>60.0</td>
<td>53.7</td>
<td></td>
</tr>
<tr>
<td>Used smartphone to track process on a health-related goal, %</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.601</td>
</tr>
<tr>
<td>Yes</td>
<td>44.5</td>
<td>48.8</td>
<td>49.8</td>
<td>48.2</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>55.5</td>
<td>51.2</td>
<td>50.2</td>
<td>51.8</td>
<td></td>
</tr>
<tr>
<td>Used smartphone to make decisions about treating a condition, %</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.474</td>
</tr>
<tr>
<td>Yes</td>
<td>40.5</td>
<td>42.4</td>
<td>40.7</td>
<td>43.3</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>59.5</td>
<td>57.6</td>
<td>59.3</td>
<td>56.7</td>
<td></td>
</tr>
<tr>
<td>Used smartphone in discussions with provider, %</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.416</td>
</tr>
<tr>
<td>Yes</td>
<td>41.3</td>
<td>39.8</td>
<td>41.2</td>
<td>39.1</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>58.7</td>
<td>60.2</td>
<td>58.8</td>
<td>60.9</td>
<td></td>
</tr>
<tr>
<td>Willing to share health data from your wearable device with health care providers, %</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.021</td>
</tr>
<tr>
<td>Yes</td>
<td>82.6</td>
<td>81.3</td>
<td>87.6</td>
<td>78.0</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>17.4</td>
<td>18.7</td>
<td>12.4</td>
<td>22.0</td>
<td></td>
</tr>
<tr>
<td>Willing to share health data from your wearable device with friends, %</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.897</td>
</tr>
<tr>
<td>Yes</td>
<td>56.1</td>
<td>57.7</td>
<td>57.2</td>
<td>57.9</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>43.9</td>
<td>42.3</td>
<td>42.8</td>
<td>42.1</td>
<td></td>
</tr>
<tr>
<td>Shared health data with health professional, %</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.977</td>
</tr>
<tr>
<td>Yes</td>
<td>14.6</td>
<td>14.2</td>
<td>14.2</td>
<td>14.2</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>85.4</td>
<td>85.8</td>
<td>85.8</td>
<td>85.8</td>
<td></td>
</tr>
<tr>
<td>Not applicable</td>
<td>8.8</td>
<td>7.2</td>
<td>7.4</td>
<td>7.1</td>
<td></td>
</tr>
<tr>
<td>Used the internet to visit social networking site, %</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.075</td>
</tr>
<tr>
<td>Yes</td>
<td>66.9</td>
<td>74.3</td>
<td>71.7</td>
<td>75.7</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>33.1</td>
<td>25.7</td>
<td>28.3</td>
<td>24.3</td>
<td></td>
</tr>
<tr>
<td>Used the internet to share health information, %</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.968</td>
</tr>
<tr>
<td>Yes</td>
<td>13.0</td>
<td>14.2</td>
<td>14.2</td>
<td>14.2</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>87.0</td>
<td>85.8</td>
<td>85.8</td>
<td>85.8</td>
<td></td>
</tr>
<tr>
<td>Used the internet to watch health videos, %</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.080</td>
</tr>
<tr>
<td>Yes</td>
<td>37.2</td>
<td>40.8</td>
<td>37.4</td>
<td>42.6</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>62.8</td>
<td>59.2</td>
<td>62.6</td>
<td>57.4</td>
<td></td>
</tr>
</tbody>
</table>

Table 1 shows the characteristics of health information source, sharing, and social networking service use. Post-pandemic respondents were significantly less likely to use electronics to look for health information (P=0.029). After the pandemic, people were more likely to use a computer, smartphone, or other electronic means to look up medical test results (P=0.018). Surprisingly, post-pandemic respondents were significantly less willing to share health data from wearable devices with their health care providers (P=0.021), while more respondents shared health information with friends. The post-pandemic respondents were more likely to visit social networking sites.

**Discussion**

The context in which patients consume health information has changed dramatically with the diffusion of the Internet, advances in telemedicine, and changes in media health coverage. Digital communication technologies are playing an important role during the COVID-19 pandemic. There is a statistically significant difference in the increased number of sources of PGHD and the use of social networking services after the pandemic. Participants were much more likely to use credible sources and health information sources, especially if they were older, more educated, and had higher literacy levels. The post-pandemic participant group had a much heavier reliance on sources that are often engaging with them in an active information-sharing manner. Leveraging digital platforms and population informatics such as mobile health and social media along with PGHD could offer unique opportunities to develop effective self-monitoring and management strategies for the new normal of the pandemic.

**References**

Comparing the Phenomic Profile of All of Us Research Program and National COVID Cohort Collaborative

Kyle P. Webb, M.S. 1, David J. Schlueter, Ph.D. 1, Jacob Keaton Ph.D. 1, Tracey Ferrara, Ph.D. 1, Ariel Williams, Ph.D. 1, Joshua C. Denny, M.D., M.S. 1
1 NHGRI/NIH, Bethesda, MD

Introduction
The National COVID Cohort Collaborative (N3C) and the National Institutes of Health (NIH) All of Us Research Program provide researchers with a wealth of observational data to assist their goals in understanding disease progression and treatment. N3C consists of more than 3.7 million patient electronic health records (EHRs) from over 45 institutions who were tested for COVID-19, exhibited symptoms consistent with COVID-19, or individuals infected with pathogens such as SARS 1, MERS and H1N1. The All of Us Research Program (AoU) is a disease-neutral longitudinal cohort study that seeks to enroll one million participants and spans a variety of data including EHR, biometric, survey, and whole genomes. As of March 10, 2021, AoU contains 315,298 participants, of whom 190,858 have condition occurrence EHR data available in the Research Workbench. AoU collects all available EHRs for individuals, with some providing over 30 years of EHR data.

The goal of this research was to investigate common phenotypic patterns associated with COVID-19 between these data enclaves. Condition occurrence terms were mapped to phecodes1, 2 which aggregate ICD9CM and ICD10CM codes into 1,817 related diagnostic groups and are commonly used in Phenome Wide Association Studies (PheWAS). In addition to directly comparing phecode frequencies from AoU and N3C, we also stratified the data by demographic information and COVID-19 diagnosis (only for those in N3C, since AoU does not provide this data). We provide possible reasons to explain differences in rates of occurrence for phecodes across these substrata.

Methodology
AoU participants were matched by the age distribution for those in N3C and had limited EHR history to four years (since the earliest condition start date in N3C is July 19, 2016). Counts of the unique number of patients were tabulated for each phecode and those with less than 20 were omitted. Log odds ratios were computed where the numerator defined the odds of the phecode’s occurrence in N3C and the denominator defined the odds of the phecode’s occurrence in AoU. With this formulation, a larger value in the log odds ratio compares to a more common frequency of that phecode in N3C than AoU and vice-versa. Chi-square tests for significance were performed and this process was repeated for phecode samples stratified by gender, race, ethnicity, and age groupings.

Results
The number of people in the N3C sample with COVID-19 and without COVID-19 were 367,029 and 1,768,684 and the AoU sample included 76,789 participants. The dashed line in Figure 1 shows the Bonferroni threshold (log Chi-
square statistic of 2.85 for a p-value of 3.201*10^-6) and any phecode with an absolute log odds ratio greater than 1.9 or a log Chi-square statistic greater than 8.5 is labeled in black. Phecodes with a log Chi-square statistic less than zero were omitted. Table 1 provides the counts and percentages for the labeled phecodes in Figure 1 and Table 2 provides phecode counts after filtering on log odds ratios for phecodes and p-values less than the Bonferroni threshold.

Table 1. Corresponding table of participant counts and percentages (in italics) for labeled phecodes from Figure 1

<table>
<thead>
<tr>
<th>Phecode</th>
<th>Description</th>
<th>N3C Covid+ N (%)</th>
<th>N3C Covid- N (%)</th>
<th>AoU N (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>480.2</td>
<td>Viral pneumonia</td>
<td>32,753 (8.92)</td>
<td>13,450 (0.76)</td>
<td>294 (0.38)</td>
</tr>
<tr>
<td>350.6</td>
<td>Disturbances of sensation of smell and taste</td>
<td>13,010 (3.54)</td>
<td>16,509 (0.93)</td>
<td>389 (0.51)</td>
</tr>
<tr>
<td>509.2</td>
<td>Respiratory insufficiency</td>
<td>5,920 (1.61)</td>
<td>5,968 (0.34)</td>
<td>81 (0.11)</td>
</tr>
<tr>
<td>316</td>
<td>Substance addiction and disorders</td>
<td>9,334 (2.54)</td>
<td>95,831 (5.42)</td>
<td>7,938 (10.34)</td>
</tr>
<tr>
<td>850</td>
<td>Hemorrhage or hematoma complicating a procedure</td>
<td>122 (0.03)</td>
<td>1,327 (0.08)</td>
<td>318 (0.41)</td>
</tr>
<tr>
<td>689</td>
<td>Disorder of skin and subcutaneous tissue NOS</td>
<td>1,430 (0.39)</td>
<td>10,613 (0.6)</td>
<td>3,990 (5.2)</td>
</tr>
<tr>
<td>386</td>
<td>Vertiginous syndromes and other disorders of vestibular system</td>
<td>137 (0.04)</td>
<td>1,208 (0.07)</td>
<td>237 (0.31)</td>
</tr>
<tr>
<td>521.2</td>
<td>Dental abrasion, erosion and attrition</td>
<td>26 (0.01)</td>
<td>299 (0.02)</td>
<td>83 (0.11)</td>
</tr>
<tr>
<td>149.4</td>
<td>Cancer of larynx</td>
<td>8,328 (2.27)</td>
<td>20,650 (1.17)</td>
<td>62 (0.08)</td>
</tr>
<tr>
<td>521</td>
<td>Diseases of hard tissues of teeth</td>
<td>136 (0.04)</td>
<td>1,076 (0.06)</td>
<td>247 (0.32)</td>
</tr>
<tr>
<td>117.2</td>
<td>Coccidioidomycosis</td>
<td>39 (0.01)</td>
<td>310 (0.02)</td>
<td>113 (0.15)</td>
</tr>
<tr>
<td>520.1</td>
<td>Hereditary disturbances in tooth structure</td>
<td>87 (0.02)</td>
<td>619 (0.03)</td>
<td>146 (0.19)</td>
</tr>
<tr>
<td>611.3</td>
<td>Lump or mass in breast</td>
<td>575 (0.16)</td>
<td>3,282 (0.19)</td>
<td>1,311 (1.71)</td>
</tr>
<tr>
<td>783</td>
<td>Fever of unknown origin</td>
<td>65,708 (17.9)</td>
<td>303,246 (17.2)</td>
<td>4,998 (65.1)</td>
</tr>
</tbody>
</table>

Table 2. Counts of phecodes (with p-values < Bonferroni threshold) by COVID groups for filters on log odds ratios

<table>
<thead>
<tr>
<th>Log Odds Ratio (LOR) Filter</th>
<th>N3C Covid+</th>
<th>N3C Covid-</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>1127</td>
<td>940</td>
</tr>
<tr>
<td>LOR &gt; 0</td>
<td>222</td>
<td>660</td>
</tr>
<tr>
<td>LOR &gt;= 1.25</td>
<td>11</td>
<td>16</td>
</tr>
<tr>
<td>0.8 &lt;= LOR &lt; 1.25</td>
<td>21</td>
<td>39</td>
</tr>
<tr>
<td>LOR &lt; 0.8</td>
<td>1,095</td>
<td>885</td>
</tr>
</tbody>
</table>

Discussion

Overall, the N3C COVID control group and the AoU sample presented more phecodes with a higher rate of occurrence than the N3C COVID positive group. Significant findings from both analyses highlight medical conditions associated with COVID-19 (e.g. phecodes 480.2, 350.6, and 509.2 shown in Table 1). Of note, in this analysis, we found it important to match individuals by length of EHR record, since individuals in AoU tended to have much longer EHR records than in N3C. Comparison studies, such as these, provide insight into the populations and outcomes when using these cohorts for research.

References

On Constraints and Considerations for Extending Support for Natural Language Processing-Based FHIR Resource Generation

Andrew Wen, MS¹, Luke V. Rasmussen, MS², Daniel Stone, BS¹, Sijia Liu, PhD¹, Prakash Adekkanattu, PhD³, Pascal S. Brandt, MSc⁴, Jennifer A. Pacheco, MS², Yuan Luo, PhD², Fei Wang, PhD³, Jyotishman Pathak, PhD³, Hongfang Liu, PhD¹, Guoqian Jiang, MD, PhD¹

¹Mayo Clinic, Rochester, MN; ²Northwestern University, Chicago, IL; ³Weill Cornell Medicine, New York, NY; ⁴University of Washington, Seattle, WA.

Introduction
With an estimated 80% of clinically relevant information being stored in unstructured form, natural language processing (NLP) has been of significant interest in the digital healthcare community, particularly with respect to development of digitized clinical decision support and learning health systems. Concurrently to these efforts, interest in deploying these same systems outside of academic environments and into clinical settings has highlighted issues relating to algorithm portability and data interoperability, leading to a push to adopt standards for health information exchange. Most notably, the HL7 Fast Healthcare Interoperability Resources (FHIR) standard has seen significant adoption amongst the healthcare community, with many of the dominant EHR solutions, including Epic and Cerner, offering FHIR-compliant representations of their data. Despite this, these solutions largely focus on presenting FHIR-based representations of information already encoded in structured format and do little to present FHIR-based views of unstructured information, such as that which can be obtained through NLP. This hinders portability, deployment, and adoption of any applications that rely on NLP as an information extraction component, and efforts to circumvent this issue are problematic as the composition of what information is contained within structured vs. unstructured data varies between institutions.

It is therefore highly desirable to also standardize NLP artifacts to the FHIR standard. To that end, in a prior study, we introduced the NLP2FHIR pipeline¹, which aimed to address this deficiency by offering generalized clinical NLP information extraction capability with FHIR-resource output. Despite this effort being successful, there are several considerations to be made with respect to FHIR resources generated from NLP output that may cause the generated output to have semantically relevant differences as compared to if that same output were to be generated from structured data. To that end, we have proposed several FHIR-compliant NLP-specific extensions for inclusion with any NLP-produced FHIR resource¹. Here, we discuss our experience implementing mappings for NLP-specific FHIR resource extensions, and present lessons learned from our experience.

Methods

```
"extension": [
    {
        "url": "https://raw.githubusercontent.com/BD2KOnFHIR/NLP2FHIR/master/StructureDefinition/conditional-modifier",
        "valueBoolean": false
    },
    {
        "url": "https://raw.githubusercontent.com/BD2KOnFHIR/NLP2FHIR/master/StructureDefinition/negated-modifier",
        "valueBoolean": true
    },
    {
        "url": "https://raw.githubusercontent.com/BD2KOnFHIR/NLP2FHIR/master/StructureDefinition/certainty-modifier",
        "valueBoolean": false
    },
    {
        "url": "https://raw.githubusercontent.com/BD2KOnFHIR/NLP2FHIR/master/StructureDefinition/is-document-date",
        "valueBoolean": true
    },...
]
```

*Figure 1.* (Truncated) Example of a populated extension definition for an NLP-derived FHIR resource derived from clinical text. Here, we see specifically the conditional, negated, and certainty modifiers, as well as a is-document-date modifier.
NLP2FHIR version 1.0.0 (https://github.com/BD2KOnFHIR/NLP2FHIR) was used as a baseline for modification, while the NLP extensions proposed in Hong et al. were used as a base guideline for implementation. FHIR output produced by the NLP2FHIR pipeline was then incorporated into a variety of downstream pipelines including obesity/comorbidity detection and colorectal cancer modeling. Data representation and modeling issues encountered during implementation and further usage in downstream applications were documented and are now reported here for informative purposes. Figure 1 shows an extension field of an NLP-derived FHIR resource with the NLP-specific extensions populated.

Results and Discussion
We encountered some issues during implementation and downstream consumption of NLP-based FHIR artifacts. Firstly, an important note with NLP-extracted clinical output is the issue of temporal information: while the date a clinical document was created can serve as a general temporal indicator for extracted NLP artifacts, this is not always accurate. For instance, documents may refer to a “past history of” a condition with an unspecified date, and it would therefore not be accurate to present the condition resource’s date as the document date. On the other hand, documents may refer to conditions with a phrase similar to “patient presented on May 1st 1999 with condition x”, in which an explicit date is made known, or present a relative date (e.g. “two weeks ago”) to the document, from which an explicit date can be derived. Such information would be appropriate for inclusion in the resource, but it is not feasible to completely resolve this issue and harmonize NLP-based output with that of structured data, as the requisite information is simply not present. Rather, it is important to make note of exactly what kind of date (document date vs. an explicit date) is used in conjunction with the resource, and downstream consuming applications should be aware of this distinction if applicable to their use case. In addition to the “term_temporal” modifier proposed in Hong’s original work, we therefore also propose a “is_document_date” true/false extension flag that denotes this difference.

Another issue encountered when implementing NLP extensions was value set consistency, which is closely tied to the main reason why one would go through standardization in the first place. The certainty modifier extension and other contextual modifiers are prime examples of this issue: unlike the semantic equivalent in the cTAKES NLP system, which offers only Boolean output for the assertion modifier, the ConText algorithm in use by NLP2FHIR for these components offers more nuanced options, with not only the standard “certain” and “possible”, but also “hypothetical”, which is not semantically equivalent to “possible” but nevertheless does contain some clinical meaning. On the other hand, the OHDSI CDM standard defines uncertainty as both “true/false” or “high/moderate/low”. As we are discussing these NLP extensions in the context of standards, it is important that these definitions and the way they are standardized into FHIR resources be consistent even across differing NLP systems, such that consuming applications can expect that the way the information is presented does not change. For that reason, while we have adopted a boolean flag for such cases to normalize to for our particular use cases, further discussion amongst the community is necessary.

Finally, it should be noted that much like structured data FHIR representations, NLP-derived FHIR output can suffer from non-normative differences in levels of granularity. NLP2FHIR is by default greedy-matching in behavior, meaning that it will take the maximum length match for normalization. A “ST-Elevated Myocardial Infarction” will therefore be codified using the SNOMED CT code specific to that condition, rather than just “Myocardial Infarction” in text. Depending on the structured data source, this behavior may be inconsistent and/or the supplementary information may be presented in a different manner when compared to structured data. Consuming applications should be aware of this distinction, and handle appropriately if necessary.

References
Predicting COVID-19 Regional Case Loads with Synthetic Data

Adam Wilcox PhD, Noa Zamstein PhD, Randi Foraker PhD MA, Jason A. Thomas BS, Kenneth J. Wilkins PhD, Jon D. Morrow MD MA MBA

aUniversity of Washington School of Medicine, Seattle, WA; bMDClone Ltd., Be’er Sheva, Israel; cWashington University in St. Louis, School of Medicine, St. Louis, MO; dNational Institutes of Health, Bethesda, MD; eNew York University School of Medicine, New York, NY

Introduction

One of the major challenges for researchers, public health officials and policy makers during the COVID-19 pandemic in the United States (US) has been predicting future disease trajectories in the population based on current infections, hospitalizations and deaths. Public health officials, policy makers and health care leaders have many different options to address infection rates in the population with varying impact on the disease and economies, ranging from mask mandates, restricting elective procedures in hospitals, to lockdowns. Using publicly-available data on infections, hospitalizations and deaths, many researchers have attempted to predict regional and local disease trajectories with limited success. One reason regional prediction has been difficult is that varying levels of community testing and hospital resources can lead to differences in detection, creating regional inconsistencies in public infection data (1). Data from emergency departments (EDs) on the number of positive cases may be a more accurate indicator of current community disease activity, which may be better for predicting future case levels; however, ED case levels are not shared in public data sets and are generally only available at the institutions providing care. The National COVID Cohort Collaborative (N3C) brings together patient-level data across multiple institutions in different geographic regions in a centralized repository; these data include positive test results and site of care information for patients in participating institutions. As such, N3C data create an opportunity to measure community trends in infections at different points in healthcare, such as ED visits. In this study, we use data from N3C to predict infection rates and disease trajectories regionally using overall community infection rates and ED visit rates. Because application of prediction using healthcare data at different institutions requires sharing of data, we performed this study using synthetic N3C data, which can allow broader sharing of data outside institutions while protecting patient privacy (4).

Methods

We predicted daily infection rates using prior infection data for multiple regions during regional disease “waves” of the COVID pandemic in the US. Our cohort included all patients in the N3C enclave who were lab-confirmed COVID-positive during 2020 (n=299,755). We extracted diagnosis date, associated encounter type (inpatient admission, ED visit, other outpatient visit, or unknown), state of residence, and length of stay. We used a Delayed Elasticiy Method (DEM) modeling approach (5) to forecast deaths at the state level, using cumulative cases at an earlier time (7 to 9 days, using the first 90 days for training the model). We report our results by state, highlighting Massachusetts (MA) and North Carolina (NC) as important examples. Synthetic derivative data sets were computed from the limited data sets using the MDClone ADAMS Synthetic Data Platform [MDClone Ltd., Be’er Sheva, Israel].

Results

Data from MA and NC showed distinct waves during autumn of 2020 and early in 2021, within which we were able to perform the prediction. Figure 1 shows the actual daily incidence rates from the data set that were used -- we used the first 90 days for training the model (light blue dots) and the following 90 days for prediction (dark blue dots). The 90 day period was selected from an established trough through a rising period in the epidemic curve; it was not selected to match the peak of the curve, as predicting the change in slope and the peak was a primary goal of the model. While the MA curve shows a clear inflection point and a peak in case load, the NC curve does not; the second wave in NC had not yet ended by the time these data were collected and only subsided in March 2021. Figure 2 shows the predicted cumulative cases for both states along with the actual cases, and Figure 3 shows the epidemic curves calculated as the change in the cumulative cases. Our model correctly predicted the change in slope of the epidemic curves for both states.
We demonstrated successful prediction of regional disease progression based on positive cases from electronic health record data in a national registry. We specifically used ED incidence rates in the prediction models, which improved the ability of the model to predict changes in the slope of the epidemic curves. Where there was a peak in the curve (MA), our model correctly predicted the peak period, even if the cumulative number of cases were different. Where there was only a flattening of the curve (NC), the model correctly predicted this change in slope and did not predict a peak where it did not exist.

This analysis was done using synthetic data, which is significant for how it could be used generally. The COVID-19 disease progression moved regionally in waves, either according to how the disease initially spread or due to state-specific mitigation policies. Training models based on other states’ data in earlier waves is necessary to create a prediction model that can be used for another state. Since this approach requires collaborative data sharing, synthetic data allow easier sharing of data while preserving privacy. Of note, this analysis could not have been done with de-identified “safe-harbor” data, because precise dates of service, which must be removed under the safe-harbor rules, are critical for determining epidemic curves.

The ability to predict disease progression in a community is critical for decision-making in a pandemic. Different states or regions reached different levels of prevalence at different times, which led to different mitigation measures being applied. For example, during the Autumn 2020 wave, some states had infection levels high enough that hospitals became overwhelmed, and temporary field hospital structures were created. At the same time, other states had less-severe outbreaks. However, states with less-severe levels did not know whether they needed to plan for creating field hospitals, and preventative measures were often used in anticipation of higher disease levels that did not occur. Many less-extreme decisions also were taken throughout, often based on incomplete information. Using more-accurate predictive models could have dramatically improved public and institutional response to the pandemic.

References
Enabling Digital Compassion in a Technology-Enabled Healthcare Environment: a Modified eDelphi Study to Establish Professional Competencies and Health Information Technology Standards

David Wiljer, PhD1,2,3, Gillian Strudwick, RN, PhD2,3, Rebecca Charow, MSc1,2, Nelson Shen, MHA, PhD3, Allison Crawford, MD, PhD, FRCPC2,3, Sanjeev Sockalingam, MD, MHPE, FRCPC1,2,3, and Peter Rossos, MD, MBA, FRCPC, FACP1,2
1University Health Network, Toronto, Ontario, Canada; 2University of Toronto, Toronto, Ontario, Canada; 3Centre for Addiction and Mental Health, Toronto, Ontario, Canada

Introduction

Healthcare is becoming increasingly digital, introducing innovative technology-enabled ways to deliver and expand access to care. In this evolving landscape, healthcare professionals, learners and organizations may not be prepared or equipped with the knowledge, skills, and behaviors required to use these digital tools in the delivery of compassionate care; moreover, the tools themselves, may not be designed and implemented in a manner to facilitate digital compassion. The notion of digital compassion is still in its infancy; however, it will be central to healthcare in the coming decades. The objective of this study is to identify: 1) essential digital compassion competencies for health professionals; and 2) digital compassion health information technology attributes.

Methods

This study is currently in progress and uses a modified eDelphi method for consensus-building through structured group communication amongst a panel of experts to address the study objectives1,2,3. The first round consisted of multiple online sessions/focus groups instead of a questionnaire. Subsequent rounds will follow a more traditional approach where web-based surveys will be emailed to panelists1. The number of rounds will continue until consensus is met. The rounds are as follows:

I. Round 1 Ideation: brainstorm domains of digital compassion
   o List professional competencies & technical attributes developed through small group activities & large group discussions

II. Round 2 Conceptual Clarity: define and rate each domain, assess conceptual clarity, generate competencies & attributes statements
   o Revise statements rated with a median score <5 for conceptual clarity based on feedback provided

III. Round 3 Refinement: assess competency and attribute statements; refine statements
    o Include statements with an “automatic consensus” if they have a median score >4

IV. Rounds 4 & 5 Consensus: come to consensus on inclusion & exclusion of competency & attribution statements (round 5 only if necessary)
   o Include remaining statements with a median score >3 and an “acceptable consensus”

The expert panel consists of 55 individuals (patients/patient advocates and professionals) who possess deep experiential knowledge, with over 5 years of experience in health informatics or digital health, health professional education, inter-professional practice, and/or compassion research. Majority work in a large urban setting (≥1 million people; 69%), and live in Ontario (60%), followed by British Columbia (13%), then Nova Scotia (9%), and the remaining in other provinces (Alberta, Saskatchewan, Manitoba, Quebec and New Brunswick). The expert panelists work in a variety of settings including academic health sciences centers, community-based health organizations (family medicine and long-term care), post-secondary institutions, non-profit organizations, healthcare IT or informatics consulting companies, mental health and substance abuse facilities, and government/provincial health authorities. To achieve this diversity, purposive and snowball sampling strategies were used. The expert panelists were recruited using an email recruitment letter sent directly to individuals using publicly available emails addresses and through the Associated Medical Services (AMS) and their organizational partners, such as Canada Health Infoway, Digital Health Canada, and advocacy groups. In Round 1, expert panelists were asked to ‘ideate’ the concept of digital compassion as it relates to a conceptual framework, healthcare delivery, professionalism, organizational culture and technological attributes. Due to COVID-19 recommendations, the format consisted of a modified world café methodology4,5, where a progressive discussion and brainstorming was facilitated through nine online Round 1 group sessions held from Sept-Nov 2020 using Microsoft Teams. Expert panelists participated in a
single Round 1 session with a group size ranging from 4-10 panelists, having multiple options to choose from. For each session, expert panelists were provided definitions and participated in a free-writing narrative activity to level-set knowledge and structure the brainstorming discussion. The free-writing narrative activity consisted of the expert panelists taking 5 minutes to think and write about a specific time where they did or did not experience compassion in a digital environment. Expert panelists had to write in the first person, starting with “I was,” and could approach this prompt from any vantage point they wished, (e.g., as a care provider, a patient, an administrator, or researcher). Afterwards, panelists shared what they wrote, if they felt comfortable, and for those listening, they wrote keywords of their observations, thoughts or feelings in the chat. As the panelists shared, these keywords were collected and rapidly thematically sorted by members of the study team on a Miro Board®. Once all the panelists had shared their stories, the Miro Board® was screen-shared for member checking (i.e., panelists can comment and provide feedback on the emerging themes and how their keywords were sorted). These themes were then used to discuss and identify professional competencies and technology standards or attributes that would address those emerging themes. Themes from the previous Round 1 sessions were also discussed – noting any overlaps or significant differences. Each session was audio recorded, anonymized, and transcribed for thematic analysis. An inductive, constant comparative approach is currently ongoing. In place of a large group discussion at the end, a draft list of guiding principles developed from the analysis as well as digital compassion professional competencies and technological requirements will be included in Round 2 for comments and assessment in the subsequent rounds.

Results

Key findings of Round 1 discussions are currently under review using thematic analysis. Emergent themes thus far have been organized within the following domains: Digital Environment, Processes, Professional Competencies and Technology Attributes. Within the Digital Environment domain, clinical interaction, connection, enabling compassionate care, time, and a sense of trust in the system emerged as the main themes. Organizational processes, policies/procedures that would allow healthcare providers (HCP) to deliver compassionate care included: choosing the appropriate modality for communication with patients, embedding digital tool use into existing workflows, providing support and being mindful of whether compassion was considered when selecting digital tools or systems. The professional competencies domain comprised of attitudes, knowledge and skills HCP should demonstrate/have when providing digitally compassionate care. Standards for the design of technology should encompass accessibility, ease of use, layout and seamless interaction, interoperability, security, understandability and user-centered co-design.

Discussion

Evidence-informed domains from Round 1 will be used to initiate discussion for subsequent eDelphi rounds (through online surveys), which will test and refine these domains, competency statements, and technology attributes. Data analysis will be conducted after each eDelphi round, and results from engagement activities will be analyzed and discussed with expert panel members. Mean scores for each domain, attribute and competency statement will be generated from the surveys. Following our work with Round 1, we hope to integrate technology attributes identified to guide the development of evaluation criteria for health information technology, and recommendations for education/training to support the incorporation of digital compassion knowledge, skills, attitude and behaviors by trainees and health professionals (Phase 2). Phase 3 will involve development of interventions utilizing the criteria and recommendations from Phase 2, and pilot testing these interventions for feasibility.

References

Lesbian, Gay, and Bisexual Patient Perceptions of Collaborative Communication: Implications for Access to Health Information

Alicia K Williamson, BA,1 Lindsay K Brown, MHI,1 Tiffany C Veinot, PhD,1,2 Denise L Anthony, PhD1,2,3

1University of Michigan, School of Information, Ann Arbor, MI; 2University of Michigan, School of Public Health, Ann Arbor, MI; 3University of Michigan, Department of Sociology, Ann Arbor, MI

Introduction
Patient-centered care is a crucial element of high-quality health care. Patient-centered care uses active collaboration and shared decision-making (SDM) between patients and providers to tailor care to patients’ health information needs. Collaborative communication (CC) is one key component of patient-centered care; it includes building rapport between patients and providers, exchanging health information, meeting patients’ emotional needs, involving patients in SDM, and enabling patient self-management. Patient-centered approaches to communication can establish strong interpersonal relationships and facilitate information exchange. Barriers to patient-centered approaches may impede health information exchange within clinical encounters, which could impact electronic communications through technologies such as patient portals. CC during the use of electronic health records within clinical encounters can increase patients’ engagement in their own care. Other informatics considerations such as using non-judgmental language in patient portals and collecting data on things like pronouns in an affirming manner are crucial. Such patient-centered communication (PCC) approaches can positively impact patient-provider relationships, adherence to medical advice, and diagnostic accuracy.

Despite the observed benefits, PCC with lesbian, gay and bisexual (LGB) populations may be lacking. Medical training regarding patient-centered care and communication for LGB patients is often insufficient; this may leave physicians feeling unprepared to address LGB patient needs. This lack of training, combined with general cultural biases against LGB people, may lead to negative experiences, such as providers making heteronormative assumptions. LGB patients are often aware of providers’ limited competence for dealing with LGB health issues, and are less likely to seek information from a doctor as their primary source of health information. Negative relationships with providers may also influence how LGB patients disclose information to healthcare providers.

Insufficient or biased communication with providers, as well as limitations in LGB patients’ likelihood to access health information through their healthcare providers, may exacerbate existing health disparities experienced by LGB people. For instance, LGB people have higher rates of mental health concerns, STIs, and obesity when compared to heterosexual populations. Additionally, bisexual people have high rates of negative health outcomes like anxiety, depression, suicidality as well as disparities in relation to the use of healthcare. This study seeks to determine if lesbian or gay (LG) or bisexual (B) patients differ from heterosexual patients in their perceptions of provider PCC. Few nationally representative surveys have studied differences in PCC for these groups outside of the context of cancer care, and fewer have examined bisexual patients separately from LG patients.

Methods
We used data from the Health Information National Trends Survey (HINTS) 5, Cycle 4 conducted from February 24 – June 15, 2020. HINTS includes six questions about PCC, asked of respondents who had a non-emergency health care visit in the previous 12 months: managing uncertainty, responding to emotions, making decisions, fostering healing relationships, enabling self-management, and exchanging information. HINTS also provides a PCC scale measure computed as a mean of the six measures converted to a 100 point scale. Bivariable and multivariable linear regression was used to assess whether LG or bisexual patients differed from heterosexual patients, adjusting for patient binary birth sex, age, race/ethnicity, education level, self-reported health status, health insurance status, having a regular provider, and whether the survey was completed after the start of the COVID-19 pandemic. We also conducted bivariable and multivariable logistic regressions of each of the six PCC questions, with responses of always coded as 1, and responses of usually, sometimes, and never coded as zero, indicating optimal PCC.

Results
In unweighted analyses, of 2,671 respondents, 95.2% identified as heterosexual, 2.4% identified as lesbian or gay (LG), and 2.4% identified as bisexual. The average unadjusted PCC rating was 80.8, and was 81.0 for heterosexual, 81.9 for LG and 72.0 for bisexual patients. In both bivariable and fully adjusted linear regression, the PCC rating for bisexual patients was significantly lower than for heterosexual patients (Beta=-7.0 [SE 2.6] for multivariable model), but there was no significant statistical difference between heterosexual and LG patients (Beta=1.7[SE 2.6]). In multivariable logistic regressions, bisexual patients were significantly less likely than heterosexual patients to report optimal levels of PCC for responding to emotions (OR=0.58[CI 0.34-0.98]), making decisions (OR=0.60[CI 0.36-1.0]), and enabling self-management (OR=0.53[CI 0.32-0.88]).

**Discussion**

As indicated by the results, bisexual patients reported significantly lower levels of PCC compared to heterosexual patients. However, there were no significant differences between LG and heterosexual patients. These findings for bisexual patients supports literature identifying negative attitudes towards bisexual people, further demonstrating the need for special attention paid to bisexual patients and their PCC needs. Poorer PCC may result in bisexual patients accessing health information from sources other than their provider, such as the internet.

Despite increased provider training on cultural competence for LGB groups, these findings indicate that more work is needed to improve patient-provider communication, particularly for bisexual patients. Improving clinical communication between LGB patients and providers may improve care and reduce health disparities. LGB people are more likely to search for health information online, and less likely to seek information from a provider, so improving clinical relationships/interactions between LGB groups and providers may improve health information access for LGB patients. Changing perceptions of communication may also impact the likelihood of using mHealth technologies, such as patient portals, to communicate with providers. As new technologies are used increasingly in healthcare, such as patient portals and telehealth, ensuring high quality PCC becomes paramount to facilitating quality care and securing equitable treatment of bisexual patients, other sexual minorities, and indeed all patients.

**References**

Improving Clinical Decision Support by Empowering Users: The Clickbusters Program

Adam Wright PhD, Elise Russo MPH, Arianna E. Nimocks, Jon G. Jackson PharmD, Jonathan P. Wanderer MD MPhil, Neal Patel MD MPH, Kevin B. Johnson MD MS, Allison B. McCoy PhD

Vanderbilt University Medical Center, Nashville, TN

Introduction:
Clinical decision support (CDS) systems have been shown to improve quality and safety of care (1, 2); however, poorly designed CDS can frustrate users and lead to alert fatigue. Development of CDS typically involves knowledge engineers, analysts, subject matter experts (3). User feedback is also valuable for evaluating and improving CDS (4); however, users are rarely empowered to take ownership of CDS.

Recently, many organizations, including our own, have begun investing in “Physician Builder” programs, where interested clinicians can get training on electronic health record (EHR) customization and build content in their area of expertise (5). Despite the name, physician programs are often open to other healthcare professionals, such as nurses, pharmacists, advanced practice providers and informaticians.

Vanderbilt University Medical Center (VUMC) has a large and effective physician builder program with 69 participants. These builders were trained and certified to build and maintain clinical content and functions in our Epic EHR. While they had developed documentation tools, order sets, reports and CDS alerts, they did not focus on optimizing existing alerts. At the start of our project, VUMC had 419 clinical decision support alerts, many of which had not been reviewed since the EHR went live.

Methods:
To improve decision support at Vanderbilt, the Vanderbilt Clinical Informatics Center (VCLIC) developed a new program to review, refine and validate CDS alerts, called Clickbusters. Clickbusters started with Vanderbilt’s existing physician builder program; Clickbusters built on this foundation with the following components:

1. Empowering our physician builders to “adopt” one or more CDS alerts in their area of expertise, evaluate the alerts and directly make improvements, with reduced bureaucracy.
2. Creating an inventory of our CDS alerts, and prioritizing them for review based on firing rate, acceptance rate and complexity.
3. A ten step Clickbusting process, shown in the figure.
4. A series of videos and knowledge base articles which taught participants how to analyze and improve a CDS alert.
5. A management process and database to support tracking of participant progress.
6. Centralized support for participants who needed assistance.
7. Gamification, including a point system, leader board, certificates, trophies and prizes.

Results:
Prior to beginning the Clickbusters initiative, VUMC had 305,841 alert firings (11,164 of which were interruptive) each week. The alerts were placed into 141 logical groups (for example, alerts related to suicide screening were grouped together). We conducted two rounds of the Clickbusters program: one in Mar-May 2020, and the other in June-Sep 2020. In the first round, eight participants selected 18 alert groups (29 total alerts) for Clickbusting. 13 alerts were modified and 4 alerts turned off, leading to a reduction of 49,026 clicks per week (a 16% reduction). In the second round, twenty participants selected 24 alert groups (55 total alerts) for Clickbusting, resulting in 29 alerts modified, 6 alerts turned off, and 22,201 weekly clicks busted (a combined 23% reduction):

<table>
<thead>
<tr>
<th></th>
<th>Round 1</th>
<th>Round 2</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participants</td>
<td>8</td>
<td>20</td>
<td>24</td>
</tr>
<tr>
<td>Alert Groups</td>
<td>18</td>
<td>24</td>
<td>42</td>
</tr>
<tr>
<td>Total Alerts Analyzed</td>
<td>29</td>
<td>55</td>
<td>84</td>
</tr>
<tr>
<td>Alerts with no modifications needed</td>
<td>12</td>
<td>20</td>
<td>32</td>
</tr>
<tr>
<td>Alerts with Modifications</td>
<td>13</td>
<td>29</td>
<td>42</td>
</tr>
<tr>
<td>Alerts Turned Off</td>
<td>4</td>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>Weekly Clicks Busted</td>
<td>49,026</td>
<td>22,201</td>
<td>71,227</td>
</tr>
</tbody>
</table>

We gave each winner a Golden Computer Mouse Trophy, as well as an Amazon gift card ($250, $150 and $100). All received a framed certificate, and the option of having VCLIC leadership send a letter of commendation to their department chair.

Key accomplishments included elimination of a frequently-firing alert related to enrollment in a congestive heart failure readmission risk reduction program that no longer existed, targeted improvements to several medication safety-related alerts, and refinement of alerts related to bronchiolitis and to administration of live-attenuated vaccines to immunocompromised patients. The Clickbuster participants were poised to make these improvements given their personal experience with the alerts, their clinical expertise and their review of relevant data.

In addition to the direct improvements in CDS, the initiative also increased user engagement and involvement in CDS. After Clickbusters, our CDS team benefited from a new corps of users with increased interest, engagement and knowledge of CDS, who could also serve as liaisons to clinical departments as diverse as pediatrics, general medicine, oncology, cardiology, surgery, nursing and pharmacy. This program helped build a culture of continuous evaluation and improvement of clinical content in the EHR.

Conclusion:
At VUMC, the Clickbusters program was useful for improving the quality of CDS alerts and involving more users in the process of evaluating and improving CDS. The process could be readily replicated at other clinical sites, and applied to other functions of the EHR, such as order sets, clinical documentation tools and information displays.

References:
Multi-horizon prediction for extracorporeal support in COVID-19 patients

Bing Xue, Msc1, Neel Shah, MD1, Hanqing Yang1, Charles Ziegenbein1, Thomas Kannampallil, PhD1, Philip R. Payne, PhD1, Chenyang Lu, PhD1, Ahmed S. Said, MD PhD1
1Washington University in St. Louis, St. Louis, MO

Keywords: Clinical Decision Support, COVID-19, Data Analytics, Machine Learning

Introduction
The SARS-CoV-2 (COVID-19) pandemic has put unforeseen strains on the global healthcare system. These challenges are particularly pronounced when deploying high-risk, resource intensive therapies, such as extracorporeal membrane oxygenation (ECMO). ECMO is a life sustaining therapy, typically initiated only after standard therapies have been exhausted, but the required expertise and resources to provide ECMO only exist in advanced centers. Guidelines continue to discourage commissioning new ECMO centers during pandemics and encourage early recognition of patients needing ECMO to allow resource allocation within ECMO capable centers or early transfer to improve patient outcomes(1). However, there remains a lack of the ability to predict which patients will require ECMO and when, causing unexpected constraints even at high resource institutions. In response to this challenge, we evaluated a time-aware prediction approach(2), developing a machine learning based multi-horizon prediction tool, “ForecastECMO”. Our prediction model focuses on identifying patients who are likely to require ECMO at various prediction horizons, providing clinicians with a needed decision support tool.

Methods
Clinical deterioration is often preceded by a period (up to hours) of changes in vital signs or laboratory results that could be unrecognized. As patients deteriorate, there is a paralleled increment in lab frequency and interventions. To account for the variability in volume of input variables at different timepoints and to identify patients during windows of increasing instability to facilitate early interventions, we developed multi-horizon predictive models called ForecastECMO, comprised of a set of machine learning models trained to predict ECMO at different time horizons. The tool was developed by constructing the optimal machine learning model at each time horizon X (hours) to predict ECMO at X hours prior to ECMO initiation. At each clinical decision point, the set of predictive models corresponding to different prediction horizons (X values) are invoked, using clinical data available till that decision point to generate a set of predictions for the need for ECMO in the next X hours. The association of the alert with the time horizon for the prediction allows clinicians to plan ECMO intervention.

This study was approved by our institutional review board with waiver of consent. Data were extracted from a registry of all patients tested for COVID-19 at our institution since January 1, 2020. The database utilizes EHR-derived data spanning 15 hospitals part of the Barnes-Jewish Healthcare system. We included all patients with a positive COVID-19 test associated with a hospital admission for ≥ 24 hours. We excluded non-ICU admissions and those meeting institutional ECMO contraindications (age >70 years and BMI > 45 kg/m²). Patients less than 3 years old were also excluded given the difference in their normal range of vital signs compared to adults and further, because there are no reported cases of patients less than 3 years receiving ECMO support for COVID-19(3, 4). Patients who received ECMO underwent detailed chart review to confirm that ECMO was directly related to COVID-19. Variables included in the ensuing database included demographics, comorbidities, laboratory values, medications, and flowsheet variables (including vital signs and respiratory support data). Clinical interpretation was incorporated into laboratory variables; creating test indicators (0: within normal range, 1: above normal range, or -1: below normal range). For patients transferred between hospitals within our system prior to ECMO, variables from the pre-transfer encounters were concatenated into the input data. Categorical variables were converted to numeric discrete values by factorization. Continuous variables were first processed by 5-percentile outlier removal, and then normalized to range [0,1] by min-max scaling. Missingness of measurements was taken into model structure.

A set of Gradient Boosting Tree (GBT) predictive models were trained for time horizons ranging 0 - 100 hours prior to ECMO initiation (51 steps of 2 hours). To train the predictive model at a time X, the input data for ECMO positive patients included measurements collected between admission and X hours before ECMO initiation, and the input data for ECMO negative cases included the worst measurements collected during hospitalization. Models were trained and evaluated with 5 random shuffles of 5-fold stratified cross validation.

Results
The registry contained 33,017 COVID-19 positive patients from March 13, 2020 to January 7, 2021. After filtering by inclusion and exclusion criteria, we identified 67 ECMO positive and 2,251 ECMO negative COVID-19 positive...
patients, admitted to ICU, with hospital stays ≥24 hours (Age: 52.88±15.41 years, BMI: 29.08±7.21 kg/m², 56.13% males, LoS: 337.65±421.39 hours). We identified a total of 348 variables of potential interest (26 demographics and comorbidities, 124 flowsheets, 52 medications, 73 laboratory variables and 73 laboratory indicators).

The model’s performance of was compared against to an a GBT model of 30 a priori identified most clinically relevant variables (Clinical GBT) and a linear model of PaO2/FiO2 (PF ratio), the only available recommendation for ECMO decision making(5). In addition to a LR regression model of all variables and a LR model of Sequential Organ Failure Assessment (SOFA) score(6, 7). Figure 1 illustrates how the model outperformed the risk scores on both the Area Under (AU) Receiver Operating Characteristic (ROC) Curve and the AU Precision Recall Curve (PRC)(8, 9) evaluation metrics.

**Discussion**

We describe a multi-horizon machine learning model capable of predicting the need for ECMO support several hours prior to initiation. Current guidelines emphasize the value of worsening respiratory failure immediately before ECMO initiation in decision making, yet fall short in providing clinicians with early warning tools for triaging resource allocation or early patient transfer to ECMO capable centers(10). Our proposed ForecastECMO model outperformed the PRESET/SOFA-based ECMO prediction in both the AU ROC and AU PR Curves across all time ranges. The model performance in this time frame, in the absence of input variables from the expected escalation in support and monitoring in the hours preceding ECMO, emphasizes the potential clinical utility of the ForecastECMO tool in identifying patients at risk of needing ECMO support in a relevant time-frame, for clinicians to reallocate resources or transfer patients to centers with more ECMO expertise.

**Conclusion**

We have developed and validated a machine learning based multi-horizon prediction tool (ForecastECMO), which has clinical utility as an ECMO decision support tool that could be used to improve patient outcomes and relieve unexpected resource burden on healthcare systems during a pandemic or equivalent public health emergencies.

**References**

2. Li D, Lyons P, Lu C, Kollef M. Deep Alerts: Deep Learning Based Multi-horizon Alerts for Clinical Deterioration on Oncology Hospital Wards. AAAI Conference on Artificial Intelligence (AAAI-20); February 20202020.
Synthetic Data to Support Engineering and Demonstrations in the All of Us Research Program

Chao Yan, MS¹, Steve Nyemba, MS², Kelsey Ross, PhD², Ziqi Zhang, BS¹, Francis Ratsimbazafy, PhD², Bradley A. Malin, PhD¹,²
¹Vanderbilt University, Nashville, TN;
²Vanderbilt University Medical Center, Nashville, TN

Introduction

One of the greatest barriers to the advancement of biomedical research is a lack of data with sufficient diversity from patients and research participants. The All of Us Research Program¹, launched in 2018 with a goal of enrolling more than 1 million American volunteers, aims to change this situation by collecting and provisioning access to a wide range of personal data derived from, but not limited to, electronic health records (EHRs), biospecimens, questionnaires, behavior and physiological data. The general belief is that making such data available on a broad scale will facilitate hypothesis generation and, given sufficiently large quantities of data, hypothesis testing regarding a number of diseases and their treatments. While making data accessible is a core goal of the program, there is also a belief that maintaining the privacy of the participants is critical to ensuring trust and sustaining a long-lasting and healthy relationship. As such, it is the priority of the program to uphold privacy principles throughout the life cycle of the All of Us program.

At the core, All of Us is a software suite, built on top of a cloud-based platform, which provides for analytic tools that access the participant-level data. To develop and develop such a suite over time, it is necessary to provide systems developers with access to such data, so that system prototypes can be engineered and the functionality and workflows associated with such systems can be sufficiently stress-tested. Moreover, it is critical to develop multiple representative case studies, based on participant-level data, for the purpose of system demonstration to the public (as well as approved prospective researchers). To mitigate privacy risks, the program developed synthetic data to support these use cases.

Generative adversarial networks (GANs) are a recently developed technique that have shown a remarkable ability to simulate structured EHR data, through a deep neural network, with a realistic feel, while simultaneously protecting privacy². GANs are composed of a generator which is forced to produce increasingly realistic instances, such that an evolving discriminator, cannot distinguish them from real data. In this abstract, we report on how we developed and applied a GAN-based framework to support the engineering and demonstration use cases of All of Us.

Methods

We focused on a dataset of approximately 235,000 participants and their associated demographics, physical measurements, and survey responses. These data are stored across multiple database tables based on the observational medical outcome partnership (OMOP) schema³. These tables were processed by a generative adversarial network (GAN) and synthesized via an iterative and incremental process. The All of Us outreach program has a study that focuses on the analysis of participant provided information (PPI) in the form of gender, race, ethnicity and various attributes in the participants’ EHR, such as height and weight. We realized the dataset by applying filters, joins and projections across the OMOP tables. The projections capture attributes of interest and filter out unnecessary records.

The GAN training process is iterative with a goal of determining an adequate representative sample size and number of epochs that can enable the neural network to converge. Once adequate sample size is determined and epochs, we perform this task in parallel in order to have multiple candidate models. Each model is used to generate data, and the models that will be retained are the ones with the least rows copied. The GAN uses a statistical method during training and will discard outliers that can be critical to a study for instance participants with diagnosis codes for rare conditions like Wolfram syndrome. Outliers are reinserted at random according to the rate at which they are observed in the original data.

Results

The synthetic data generated using the GAN-based framework were leveraged to support gold standard software development and release management processes. Synthetic data not only enable the Researcher Workbench (RW) software engineers to develop and test new features, but also to complete critical quality control and assurance tasks.

In addition to supporting software development, synthetic data are relied upon to support All of Us RW outreach and training. Since the RW launched in May of 2020 for beta testing, the synthetic data described here have been used in...
more than 30 researcher outreach and training events with over 1,800 users. The primary audience for these outreach demonstrations are prospective researchers who may be interested in registering to become All of Us Researchers. During these demonstrations, members of the RW team walk through example analyses using synthetic data within a software environment that mirrors the production RW. A primary goal of the demonstrations is to train researchers to use the tools available within the RW and educate prospective researchers key characteristics of the All of Us dataset.

Within the production All of Us RW, we provide tutorials on real data to train registered, approved All of Us researchers how to do basic analyses on All of Us data. These include tutorials that describe how to, such as querying the database and characterizing the number of participants that have contributed multiple data types in All of Us (Figure 1A). Using synthetic data in the mirror RW environment, we run the same tutorial analysis, demonstrating similar key features within the data (Figure 1B), without requiring all attendees of the demonstration to have prior approval to access participant-level data. Similarly, we regularly conduct demonstrations that show an example end-to-end analysis examining standing height data derived from two different program sources. As one might expect, in the real data (Figure 1C) we observe that the distribution of height derived from EHRs closely resembles that derived from the program baseline assessment. While the distribution of heights using the synthetic data (Figure 1D) does not precisely mirror that of the real data, the general shape and agreement of the distribution of height data has similar features to that of the real data. In practice, audience members attending these demonstrations find the synthetic data plausible enough that the demonstrations are successful in meeting their stated goals.

Discussion and Conclusions

GAN-based data synthesis framework and the yielded synthetic data have been used in large scale to support the two important use cases in the All of Us program—system engineering and demonstration. Due to the sufficiently high quality of the generated data, software development and testing, and case study demonstrations have been efficiently expanded. The great performance of synthetic data drives us to move beyond these two use cases with a goal to develop publicly accessible synthetic datasets, which can be used for research purposes. However, there exist several challenges we would highlight as opportunities for future investigation. First, due to an extremely large feature space, the current implementation of data synthesis does not represent the relationships between features and tables. New synthesis frameworks that account for consistency in the feature space need to be developed. Second, though GANs have the potential to preserve privacy, it is unclear whether a synthetic copy of the All of Us data with a very large feature set will sufficiently maintain privacy. Further investigation is needed to evaluate residual privacy risks that may be incurred in synthetic data, such as membership attacks (e.g., recognition of records in the underlying dataset upon which the synthetic data generator is based).

References

A Continuous Crowd-sourced Challenge for Benchmarking COVID-19 Health Outcome Prediction

Yao Yan, BS\textsuperscript{1,2}, Thomas Schaffter, PhD\textsuperscript{2}, Timothy Bergquist, PhD\textsuperscript{1,2}, Thomas Yu, BS\textsuperscript{2}, Justin Prosser, BS\textsuperscript{1}, Justin Guinney, PhD\textsuperscript{2}, Sean Mooney, PhD\textsuperscript{1}

\textsuperscript{1}University of Washington, Seattle, WA, USA; \textsuperscript{2}Sage Bionetworks, Seattle, WA, USA

Introduction
The novel Coronavirus SARS-CoV-2 has caused a global pandemic. As the case counts rise and more people are hospitalized, patient-level health data becomes a viable and crucial resource for researchers to understand disease patterns and design evidence-based interventions against the disease.[1] Due to the private and sensitive nature of patient data, restrictions have been put in place for the sharing of this data. To overcome this problem, we implemented a model-to-data continuous benchmarking challenge - the COVID-19 EHR DREAM Challenge - to enable the data science community to validate clinically relevant prediction models on a private EHR dataset without ever accessing the data.[2,3] This challenge used COVID-19 patient data from University of Washington to enable citizen science and predictive model benchmarking.

Methods
We ran the COVID-19 EHR DREAM challenge as a continuous benchmarking exercise where the datasets were updated every 2-4 weeks to incorporate new patients and update existing patients’ clinical trajectory. In our COVID-19 Challenge, we asked participants to address two clinically pressing questions: Question 1 (Q1) Of patients who receive a test for COVID-19, who will test positive and Question 2 (Q2) Of patients who tested positive for COVID-19 in an outpatient setting, who is at risk for hospitalization within 21 days. We curated two sub-datasets (diagnostic Challenge Dataset and prognostic Challenge Dataset) separately for the two challenge questions for the purpose of model training and evaluation. We also had a cumulative dataset for model analysis. Between February 2020 and January 2021, the UW COVID-19 patient dataset accumulated 108,000 patients who had been tested for COVID-19. Among all patients who received a COVID-19 test, 4,980 tested positive, 3,100 tested positive during an outpatient visit and 170 were hospitalized within 21 days after testing positive during an outpatient visit.

We implemented a model-to-data approach for the COVID-19 challenge to facilitate the delivery of participant models to the challenge datasets in order to train and evaluate their predictive models. COVID-19 patient datasets were hosted in a UW on-premises server behind a UW Medicine IT provisioned firewall. Challenge participants never had direct access to the patient data; instead, they were required to build and submit Dockerized (containerized) models. A synthetic dataset was provided to the participants to help them become familiar with the format of the data and to aid in technical debugging. A submitted model underwent validation procedure on an Amazon Web Service (AWS) cloud environment and once validated, was then pulled in a UW secure environment where it was trained and then applied to patient data (the holdout set from the full patient dataset) to generate

![Figure 1. Model performance AUROC and AUPRC (box plots) and challenge participation (grey bars) for Question 1(a,b) and Question 2(c,d) weekly. Different colors are used to represent different dataset versions.](image-url)
predictions. The performance of the models were evaluated by computing the Area Under the Receiver Operating Characteristic Curve (AUROC) and Area Under the Precision Recall Curve (AUPRC). These two performance scores were returned to participants.

**Results**

By January 2021, this challenge engaged 482 participants from over 90 teams, with 26 having submitted valid models to challenge questions. Datasets for Q1 had 30 weeks with 7 versions and for Q2 had 18 weeks with 4 versions. The best scores achieved during the challenge were AUROC 0.827 and AUPRC 0.303 for COVID-19 diagnosis prediction, and AUROC 0.982 and AUPRC 0.897 for 21-day COVID-19 associated hospitalization prediction. (Challenge participation see Figure 1). In the analysis using cumulative data from different versions of the challenge dataset, the best performance for COVID-19 diagnosis prediction was AUROC 0.776 (95%CI 0.775, 0.777) and AUPRC 0.297, and for hospitalization prediction, AUROC 0.796 (95%CI 0.794, 0.798) and AUPRC 0.1875. Analysis on top 10 models submitting to Question 1 showed consistently better model performance on the female group than male group. The best performance was obtained for the 25-49 age group and the worst performance was obtained for 0-17 age group among all age groups. (Figure 2)

**Discussion**

We succeeded in implementing a model-to-data approach to operate a continuous benchmarking challenge and enable model training and evaluation on up-to-date COVID-19 patient EHR data with a worldwide data science community. The benchmarking challenge provided an unbiased evaluation of models submitted by participants. Across submitted models, we observed discrepancies of performance in this temporally evolving dataset and between demographic sub-populations (gender and age), indicating the existence of potential bias in machine learning approaches, which warranted attention prior to implementation of such models in clinical practice.

**Reference**


Lana Yeganova, PhD, Won Kim, PhD, Donald C. Comeau, PhD, W. John Wilbur, MD PhD
Zhiyong Lu, PhD.
National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM), National Institutes of Health (NIH), Bethesda, MD USA 20894

Introduction

PubMed (https://pubmed.gov) is a search engine providing access to a collection of more than 30 million biomedical abstracts. About 5 million of these articles have full text versions available in PubMed Central (PMC; https://www.ncbi.nlm.nih.gov/pmc), however, it is not currently possible for a user to simultaneously query the contents of both databases with a single integrated search.

With the growing availability of full-text articles, integrating abstracts and full texts of documents into a unified representation becomes essential for performing comprehensive biomedical literature search. An obvious benefit is improving the handling of queries that produce no retrieval in PubMed, for which incorporating full text information can yield useful retrieval results. For example, the query cd40 fmf retrieves no articles in PubMed, but finds dozens of articles in PMC, which discuss protein cd40 and a computational technique of flow microfluorometry (FMF).

Here we examine how to combine abstracts with available article full texts to improve the overall retrieval performance. We first create the PubMed Click evaluation dataset, which incorporates PubMed queries, retrieved documents and user clicks. Using the PubMed Click dataset, we establish the connection between the BM25 score (currently powering PubMed search engine (1)) of a query term appearing in a section of a full text document and the probability of that document being clicked or identified as relevant. Probability is computed using Pool Adjacent Violators (PAV) (2), an isotonic regression algorithm, providing a maximum likelihood estimate based on the observed data. Using this probabilistic transformation of BM25 scores we show an improved performance on the PubMed Click dataset, and the TREC Genomics collection (3). By using these datasets, we examine how to combine information coming from various parts of a full text document for improved retrieval. We find evidence that different sections of a full text document are of different value in deciding relevance and propose a novel method to convert the information to log odds scores which can be treated uniformly.

Methods

PubMed Click Evaluation Dataset. To address the problem, we create a new evaluation dataset based on indirect human judgements. The dataset is constructed based on retrospective analysis of PubMed queries under the assumption that a clicked document is relevant to a user issuing a query. For each query, clicked documents are collected and labeled positive. Clicks on the top rank are ignored as a precaution, as these might simply represent a user’s urge to click on something indiscriminately. Documents displayed above the clicked document are assumed to have been seen by the user and rejected and are labeled negative. Documents displayed below the lowest clicked document on the document summary page are ignored as the user may not have considered them. In addition, we only consider documents for which none of the query tokens appear in the title, to remove the click bias associated with the presence of query terms in the title. We also dropped articles where the token did not appear in the full text. As the same query string may be searched multiple times within a year, we merge the data from each query and remove from the negative set documents that also appear as positives. For the evaluation dataset, we randomly sampled 2 million unique queries from the PubMed 2017 query log.

Log_Odds BM25 Term Scoring. Here we examine how to optimally use BM25 scores coming from the abstracts and full text sections to improve retrieval performance. Following (4), we consider thirteen full text section types: Abstract, Abbreviation, Caption, Discussion, Case Study, Keyword, Conclusion, Result, Methods, Introduction, Supplement, Appendix, and Generic Section Title. This means that for a given token, for any document we have potentially thirteen different BM25 scores for that token, one from each section type.

We first define the score of a token within a section of a full-text document. For each token, we compute a BM25 score representing the relevance of the token to a paragraph of text. Since there are generally multiple paragraphs within each section of a paper, we keep the largest BM25 score for a paragraph to represent the section score for that token.
Because of the structure of full text documents, the appearance of a token in different sections makes different contributions to the relevance of the document. The same BM25 score may have a different significance depending on the section: a high score in the Results section would likely be more indicative of importance than a similar score from the Methods section of a paper. We propose converting these BM25 section scores into probabilities of a document being identified as relevant, using the PAV Algorithm (2, 5). The PAV-based probabilistic transformation allows one to directly compare the value of section scores to each other. We will refer to the PAV-based transformed scores as Log_Odds BM25.

Here we compare the raw BM25 score with Log_Odds of BM25 to learn the benefits of the proposed probabilistic transformation. Following earlier studies, we consider two flavors for Log_Odds BM25 score: one based on taking maximum score over all sections (6) and the sum over section scores (7). BM25 Abstract score is also computed as it is related to the current PubMed search system.

Results

Preliminary experiments are performed on PubMed Click Dataset and TREC Genomics collection. Two retrieval ranking approaches are compared: using the raw BM25 scores and PAV-transformed Log_Odds BM25 scores. For each approach sum of scores and max score are considered for combining section scores.

Figure 1 demonstrates our findings computed on the complete set of single tokens that appeared in the queries of PubMed Click dataset and TREC genomics datasets. In both datasets, Log_Odds BM25 Sum and Log_Odds BM25 Max demonstrate an improved performance compared to Sum and Max on raw BM25 scores. Based on these results we believe that log odds scoring is a useful approach for measuring the value of the body text.

Discussion

We studied how to integrate full text and abstract information for scoring a query token. Our experimental results demonstrate that PAV based log odds scoring is a useful way to compare the contribution of occurrences of a token in different sections of a document for predicting relevance. Raw BM25 scores are not comparable with each other for making those predictions. The same BM25 score is of different value depending on the section type in which it is found.

References

7. Hearst MA, Plaunt C. Subtopic structuring for full-length document access. SIGIR93: 16th International ACM/SIGIR '93 Conference on Research and Development in Information Retrieval; Pittsburgh PA USA1993
Comparing Automated Extraction to Manual Chart Review for COVID-Specific Research Data Abstraction: A Case Study

Andrew L. Yin*, M.D., M.B.A. 1, Winston L. Guo*, B.S. 1, Evan T. Sholle* M.S. 1, Mangala Rajan, M.B.A. 1, Laura C. Pinheiro, Ph.D., M.P.H. 1, Parag Goyal, M.D. 1, Justin Choi, M.D. 1, Mark N. Alshak, M.D. 1, Han. A. Li, M.D. 1, Graham T. Wehmeyer, M.D. 1, Mark Weiner, M.D. 1, Monika M. Safford, M.D. 1, Thomas R. Campion, Jr., Ph.D. 1, Curtis Cole M.D. 1

1Weill Cornell Medicine, New York NY

Introduction
The pressing public health threat of COVID-19 fundamentally altered the landscape of clinical research, opening an urgent need for data for clinical, research, and operational applications - a need to which many institutions responded accordingly. Our institution created a COVID Institutional Data Repository (IDR), comprised of data retrieved through manual abstraction and automated extraction from EHR systems, including medications, comorbidities, labs, symptoms, and more. This resource afforded the opportunity to explore the comparison between manually and automatically derived data, further elucidating their relative strengths and weaknesses. Manual chart abstraction is often considered the gold standard for retrospective observational research, and many variables require manual adjudication from clinically trained personnel (1). However, human reviewers are not infallible and are still less accurate in certain cases while suffering the key shortcoming of linear scaling as cohort size increases (2).

In part to address these shortcomings, studies have demonstrated that automated data extraction can perform on par with manual abstraction for certain variables (3,4). Automated extraction from the EHR can have data quality issues, due to high complexity or fragmentation of data (5). Both automated and manual techniques ultimately depend on the EHR, which is not an objective, canonical source of truth but rather an artifact with its own bias, inaccuracies, and subjectivity (6,7). We sought to compare data collected through manual chart review to the data populated via automated extraction to determine their level of concordance regarding exposure to home and inpatient medications, to assess the relative performance of the two abstraction techniques by re-reviewing discrepancies to determine which was correct, and to generate a taxonomy of discordance by classifying instances of discrepancy.

Methods
To compare data collected between methods, a query was developed in SQL to identify, for each patient, whether the automated extraction agreed with manual abstraction for administration of inpatient and home medications (queries can be found at https://github.com/wcmc-research-informatics/covid_comparison). The query was designed to align with the manual abstraction instructions (e.g. searching for medication names as opposed to RxNorm codes or sometimes having requirements for duration). For each medication, we calculated Cohen’s kappa and 95% CI to quantify agreement between data abstracted manually and automatically. Based on the query output, 10% or 20 identified discrepancies for each medication, whichever was greater, were randomly selected to be audited and reviewed by two team members (AY and WG). After initial review of a subsample of discrepancies, results were classified into categories and harmonized through discussion, which was repeated until thematic saturation. Descriptive statistics were calculated once all discrepancies (n=680) were reviewed.

Results
All 25 medications and drug classes (16 inpatient and 9 home) included in the manual abstraction were included. For inpatient medications, kappa values ranged from 0.96 to 0.0. For home medications, kappa values ranged from 0.77 to 0.18. Discrepancies were divided into three principal categories: human error, ETL/mapping error, or abstraction-query mismatch. For inpatient medications, 310 discrepancies were audited: 119 were due to human error, 46 to ETL/mapping, and 145 to abstraction-query mismatch. For home medications, 370 discrepancies were audited: 77 were due to human error, 199 to ETL/mapping, and 94 to abstraction-query mismatch.

| Table 1. Rater agreement data and all calculated statistics for inpatient and home medications |
Conclusion
This study demonstrates the viability of automated collection of inpatient medication data from the medical record, with home medications showing weaker results. Institutions with large databases should be able to employ similar methods. Although the findings and implications are broadly applicable, the query was tailored to this specific IDR; replicating this analysis will require site-specific adaptation depending on EHR data transformation and database structure. Second, although the audit process was extensive, it focused on discrepancies rather than a random audit of all results, thus possibly overlooking errors among agreeing results. Third, there may be underlying bias within the data collection methods that impacted data quality for certain populations. Fourth, this study does not attempt to evaluate the severity of errors, and some errors may be considered more critical than others. Lastly, the current work focuses on the important but specific domain of medications, which we hope to expand upon in the future.

The study results call into question the consideration of one data collection method as a “gold standard” for observational retrospective research. It is important to be rigorous and transparent about data sources and methods used, accepting that errors will occur. In certain circumstances, available resources will vary, requiring alternative methods with which to acquire and analyze data. Targeting manual and automated abstraction efforts at data domains where performance is demonstrably superior can lessen the time and effort needed and enhance data quality overall.

References
PAGER-CoV-Run: An online interactive analytical platform for COVID-19 functional genomic downstream analysis

Zongliang Yue⁴, Nishant Batra¹, Hui-Chen Hsu², John Mountz², and Jake Chen*¹
¹Informatics Institute, School of Medicine, the University of Alabama at Birmingham, Birmingham, AL 435233, USA
²Division of Clinical Immunology and Rheumatology, Department of Medicine, the University of Alabama at Birmingham, Birmingham, AL 435233, USA

Background: COVID-19 has become a global public health threat for over a year. Recently, the emergence of the SARS-CoV-2 coronavirus Delta variant has brought additional concerns due to its unusually high transmissibility and potential vaccine break-through. To interpret genomic and functional genomic effects of SARS-CoV-2 on human tissues and perform comparative analysis among various SARS-CoV-2 infected samples, we developed an online tool called “PAGER-CoV-Run”. PAGER-CoV-Run takes pathways, gene sets, and gene signature information from the PAGER-CoV database and performs advanced functional genomics data analysis online for users without the need for any coding. PAGER-CoV-Run also implements advanced visualization and network analysis features, potentially serving as a model for future functional genomics tools for target identification.

Result: As of the current implementation, PAGER-CoV-Run enables the pre-configured analysis of 629 samples covering six cell types, i.e., primary human lung epithelium (NHBE), lung alveolar (A549 cell line), the transformed lung-derived Calu-3 cells (Calu3), peripheral blood mononuclear cells, leukocytes, Human induced pluripotent stem cell-derived cardiomyocytes, and four tissues, i.e., lung, liver organoid, pancreatic organoid, and nasopharyngeal. By comparing the pathways enriched in A549 cell line, we found that pathways that are associated with the interferon-alpha/beta response, the peginterferon alpha-2a/peginterferon alpha-2b response, the OAS antiviral response, and the human immune response to tuberculosis are commonly enriched shared among RSV, HPIV3, and SARS-CoV-2 infection. Our results suggest that interferon-alpha/beta treatment in RSV and HPIV3 may have similar effects for SARS-CoV-2. By comparing the pathways enriched in SARS-CoV-2 infected NHBE and SARS-CoV-1 NHBE, we found that SARS-CoV-2 specifically alters NF-kappa B signaling pathway, suppresses HMGB1 mediated inflammation by Thrombomodulin (THBD), triggers genes involved in RUNX2-mediated cell migration, and affects lipoxin and 15-eicosatetraenoic acid biosynthesis. This case study suggests that potential drugs in targeting the pathways above may further suppress SARS-CoV-2–mediated lung damage and inflammation.

Conclusion: PAGER-CoV-Run is a powerful public tool to perform functional genomic analysis across cell types and tissues. Additional data configured into the PAGER-CoV-Run should improve end-user exploration of major pathogenic mechanisms in COVID-19 patients with the potential to identify pathways that promote long-haul sides effects in various tissues. PAGER-CoV-Run is publicly accessible from http://discovery.informatics.uab.edu/pager-cov-run.
Clinical Term Embeddings from SNOMED CT

Fuad A. Zahra and Rohit J. Kate
Department of Computer Science
University of Wisconsin-Milwaukee, Milwaukee, WI, 53211

Introduction: Word embeddings represent semantic properties of words in the form of fixed-length vectors and are critical for applying deep learning based methods for natural language processing tasks, including in the biomedical domain [1]. Word embeddings are commonly obtained from large text corpora based on the assumption that words with similar meanings will appear in similar contexts. However, semantic properties of words or terms are often directly encoded in their ontologies in the form of relations between them. Examples of such ontologies include WordNet [2] in the general domain and SNOMED CT [3] in the clinical domain. For example, one can infer from SNOMED CT relations that “asthma” and “pneumonia” are similar terms because both are disorders related to lungs. Given that clinical concepts are defined in SNOMED CT in terms of their relations with other concepts (this is unlike the relations in UMLS [4]), it makes it an invaluable resource for obtaining clinical term embeddings. But SNOMED CT has remained mostly underutilized as a resource for obtaining clinical term embeddings. In this work, we used only SNOMED CT relations to obtain embeddings for clinical terms and evaluated them using four benchmark datasets of clinical term similarity. Our results show that our SNOMED CT-based embeddings performed competitively against four other embeddings that were obtained using corpus-based methods. To the best of our knowledge, there is only one previous work that had reported embeddings obtained from SNOMED CT [5], but the authors had used a different method and had evaluated it only for concept similarity and not for clinical term similarity.

Methods: We adapted the method that had been previously used to obtain word embeddings from WordNet in the general domain and was shown to perform well [6]. The method is based on the intuition that two concepts are semantically closer if in their ontological graph there are larger number of paths between them as well as if these paths are shorter. Computationally, the method proceeds as follows. We first obtain an n x n adjacency matrix M from the SNOMED CT ontological graph G where n is the total number of concepts in SNOMED CT. An entry $M_{ij}$ in this matrix is 1 if there is a relation between the concept represented by the i’th row with the concept represented by the j’th column, otherwise the entry is 0. An entry in the k’th power of matrix $M^k$ then represents the number of paths of length k between the two concepts corresponding to the matrix entry. Thus, the matrix sum $M_0 = I + \alpha M + \alpha^2 M^2 + \alpha^3 M^3 + \ldots$ where I is the identity matrix and $\alpha$ (< 1) is a decay factor, depicts the total number of paths of all lengths between any two concepts where longer paths are progressively counted with less weights. This sum can be analytically obtained using the inverse matrix operation: $M_0 = (I - \alpha M)^{-1}$.

The matrix $M_0$ thus obtained represents each concept with a vector that has as many dimensions as the number of concepts in SNOMED CT, which is more than 400,000. However, in order to be practical, an embedding needs to be of much lower dimensions. Hence, we next apply Principal Component Analysis (PCA), a well-known method for dimension reduction. Because the matrix is too large, we used Incremental PCA (IPCA) method [7] instead. We chose 850 as the dimensions for the embeddings and $\alpha = 0.75$ based on pilot studies. It should be noted that this method generates embeddings for SNOMED CT concepts (i.e. an embedding for each SNOMED CT concept ID) and not for clinical terms or words. However, for natural language processing applications as well as for our evaluation of clinical term similarity in this work, we need embeddings for clinical terms instead. Although clinical concepts in SNOMED CT are paired with their associated clinical terms, getting embeddings for clinical terms from embeddings of clinical concepts is not trivial because there could be multiple concepts or no concept in SNOMED CT corresponding to a clinical term. Hence, we use the following method to derive embeddings of clinical terms from the embeddings of clinical concepts. If a clinical term is not ambiguous, i.e. if there is a unique concept corresponding to it in SNOMED CT, then the clinical term’s embedding is considered same as the concept’s embedding. If multiple concepts correspond to a clinical term in SNOMED CT, then its embedding is obtained by vector summing all those concept embeddings to represent an aggregated meaning. In case a clinical term does not correspond to any concept in SNOMED CT, then its embedding is obtained by vector summing the embeddings of all the concepts whose clinical
terms include that term as a sub-term. This happens for less than 5% of the clinical terms in our data. In the above procedure, we first expand the synonym sets of clinical terms in SNOMED CT using more synonyms from UMLS.

Results and Discussion: We evaluated our clinical term embeddings obtained from SNOMED CT on four benchmark datasets of clinical term similarity that have been widely used previously [8]. Each of these datasets contains a list of clinical term pairs along with expert-judged scores of similarity between them. Using embeddings, the similarity between a pair of clinical terms is determined as the cosine similarity between the embeddings of the clinical terms.

In the results shown in Table 1, we report the Pearson correlation coefficient scores between the expert-judged similarity scores and the similarity scores obtained using the embeddings. We compare our results with the results obtained using four embeddings obtained using corpus-based methods as reported in [8]. The embeddings GloVe and Google News were obtained from general domain text corpora, the embedding EHR was obtained from a corpus of clinical notes, and the embedding MedLit was obtained from a corpus of biomedical articles.

It can be observed from Table 1 that our SNOMED CT embeddings obtained the best score on the first dataset and was competitive on the rest of the datasets. It was consistently much better than GloVe and Google News embeddings showing that our method gleaned far superior knowledge from SNOMED CT to encode in the embeddings than the knowledge about clinical terms the corpus-based methods could glean from general domain text. Compared to EHR, our embeddings did worse on three datasets and better on one dataset. Our embeddings did better than MedLit embeddings on three datasets and was slightly behind on one dataset. These results show that embeddings obtained from SNOMED CT without using any text corpora are comparable to embeddings obtained using state-of-the-art corpus-based methods from biomedical text corpora.

Table 1. Performance comparison of our SNOMED CT-based embeddings and four corpus-based embeddings as reported in [8] on the four clinical term similarity benchmark datasets. The numbers are Pearson correlation coefficients between the expert-judged similarity scores and the similarity scores obtained using the embeddings.

<table>
<thead>
<tr>
<th></th>
<th>SNOMED CT</th>
<th>EHR</th>
<th>MedLit</th>
<th>GloVe</th>
<th>Google News</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hliaoutakis’s</td>
<td>0.614</td>
<td>0.482</td>
<td>0.311</td>
<td>0.247</td>
<td>0.243</td>
</tr>
<tr>
<td>MayoSRS</td>
<td>0.319</td>
<td>0.412</td>
<td>0.300</td>
<td>0.082</td>
<td>0.084</td>
</tr>
<tr>
<td>Pedersen’s</td>
<td>0.547</td>
<td>0.632</td>
<td>0.569</td>
<td>0.403</td>
<td>0.357</td>
</tr>
<tr>
<td>UMNSRS</td>
<td>0.415</td>
<td>0.440</td>
<td>0.404</td>
<td>0.177</td>
<td>0.154</td>
</tr>
</tbody>
</table>

Conclusions: We presented a method to obtain embeddings for clinical terms directly from SNOMED CT ontology without using any text corpora. The embeddings we obtained performed well on four benchmark datasets compared to embeddings obtained using corpus-based methods. Developing a method that combines the two types of embeddings thus combining their strengths will be an avenue for future research.

References
Early Detect COVID-19 Presenting Symptoms and Characteristics Using Natural Language Processing on Electronic Health Records

Juan Zhao1, Monika E Grabowska2, Vern Eric Kerchberger1,1, Joshua C. Smith1, H. Nur Eken3, QiPing Feng4, Josh F. Peterson1, S. Trent Rosenbloom1, Kevin B. Johnson1,4, Wei-Qi Wei1,6

1Department of Biomedical Informatics, 2Medical Scientist Training Program, 3Department of Medicine, Division of Allergy, Pulmonary & Critical Care Medicine, 4Division of Clinical Pharmacology, Department of Medicine, 4Department of Pediatrics, Vanderbilt University Medical Center, Nashville, TN; 5Vanderbilt University School of Medicine, Nashville, TN

Introduction

The COVID-19 pandemic is a rapidly evolving public health and economic crisis, with over 34 million people infected in the United States and 194 million globally. The virus's high transmissibility, lack of native immunity, high mutability, and the dearth of effective treatments make managing COVID-19 uniquely challenging. The public health authorities first had to identify and draw attention to signs and symptoms of specific to COVID-19, which can enable rapid symptom screening, diagnostic testing, and contact tracing. Previously, the primary efforts to track COVID-19 symptoms relied on scanning scientific publications or Twitter1, deploying questionnaires2, or releasing apps to self-report symptoms, but results from publications and questionnaires are often delayed; data from scientific publications or Twitter are often delayed; data from publications and questionnaires are often delayed; data from social media or self-reported apps do not always include proper controls and lack physiological assessments to determine COVID-19 status. To timely recognize the emerging symptoms and clinical characteristics of COVID-19, we utilized natural language processing (NLP) to analyze clinical notes from Electronic Health Records (EHR) of people tested for COVID-19. We developed a method called the concept-wide association study (ConceptWAS)3 to identify clinical concepts associated with COVID-19 and performed a temporal analysis. By analyzing clinical notes of 19,692 adult patients who received COVID-19 testing at Vanderbilt University Medical Center (VUMC) between March 8 and May 27, 2020, we were able to detect loss of smell and loss of taste three weeks prior to their inclusion as symptoms of the disease by the Centers for Disease Control and Prevention (CDC).

Materials and Methods

Study setting. The study was performed at VUMC, one of the largest primary care and referral health systems serving over one million patients annually from middle Tennessee and the Southeast United States. We used data from patients represented in the VUMC EHR aged ≥18 years. The study was approved by the VUMC Institutional Review Board (IRB #200512).

Cohort definition. We identified patients who received at least one SARS-CoV-2 polymerase chain reaction (PCR) test between March 8 (when the first COVID-19 case emerged at VUMC) and May 27, 2020 (Figure A.1). The COVID-19 status was determined using the PCR test result. The case group (COVID-19-positive) was defined as patients who had ≥1 PCR positive result, and the control group (COVID-19-negative) consisted of patients with only negative PCR tests. We excluded patients who had no clinical notes on the day when the PCR test was ordered.

Data collection. We extracted clinical notes from 24 hours prior to PCR testing date (day0) for the cohort. If a patient first tested negative and then subsequently tested positive or if a patient tested positive more than once, we used the date of the first positive PCR test as day0. We also segmented the study period into a 2-week interval window and performed a temporal analysis using every 2-week cumulative data. The primary types of clinical notes that we extracted included progress notes, problem lists, Emergency Department (ED) provider notes, ED triage notes, imaging reports, social histories, etc.

NLP processing. We utilized an NLP pipeline to preprocess, analyze, extract, and identify the clinical concepts from the clinical notes. The NLP pipeline employs knowledge map concept indexer and several rule-based parsers to map from text mentions to Concept Unique Identifiers in the Unified Medical Language System (UMLS) and SNOMED Clinical Terms. We also employed additional UMLS semantic types to accommodate relevant items such as symptoms, mental or behavior dysfunction, or social behaviors. We enabled the negation detection and developed multiple rules to enhance such detection and filter out unrelated concepts, e.g., excluded any concepts that arose from family history sections, removed any sentences with future tense or subjunctive mood (e.g., "should", "could", or "if") that describe uncertainty.

Statistical analysis. We applied Firth's logistic regression to examine the association for each concept, adjusted by age, gender, and race. We chose Firth's logistic regression because it has become a standard approach for analyzing binary outcomes with small sample. We used a Bonferroni correction for the significance level. For each concept, we report the odds ratio (OR), p-values, and the prevalence in case and control groups.

Two medical students performed a manual chart review to evaluate the clinical plausibility of identified signals. For concepts p-value met Bonferroni-corrected significance, we randomly selected 10-20 notes for each concept to review whether the notes mentioned the symptoms in the expected attribute (e.g., affirmative or negated).

Results
We identified 19,692 patients with COVID-19 PCR test results during the study period. A total of 1,483 (7.5%) patients tested positive for COVID-19. Patients' mean age was 45 (44.6 ± 16.9) years. The COVID-19-positive group was younger (41.5 ± 16.2 vs. 44.9 ± 16.9), more often male (48.0% vs. 41.7%), less often white (49.6% vs. 66.7%), and newer to VUMC (EHR length 7.3 years ± 8.1 vs. 9.2 ± 8.5) compared to COVID-19-negative patients. We processed 87,753 notes from 19,692 patients. After using the NLP pipeline to process the notes, we recognized 19,595 unique concepts (including negated status). Within that corpus, we identified concepts positively associated with COVID-19 including "anosmia" (loss of smell, odds ratio [OR] = 4.97, 95% confidence interval [CI] = 3.21–7.50), "fever" (OR = 1.43, 95% CI = 1.28–1.59), "cough with fever" (OR = 2.29, 95% CI = 1.75–2.96), "ageusia" (loss of taste, OR = 5.18, 95% CI = 3.02–8.58), as well as those concepts that were negatively correlated with COVID-19 such as "depression and anxiety" and "smoking" (Fig. 1). Earlier epidemiological studies found that fewer smokers are among COVID-19 patients or hospitalized COVID-19 patients, which are consistent with our findings of the negative correlation between smoking and COVID-19.

Notably, by performing temporal analysis on every two-week cumulative data, this study detected the signal of loss of smell and taste as early as April 5, 2020 (Fig. 2), nearly three weeks earlier than the date that they were listed as COVID-19 symptoms by CDC. The chart review finally reviewed 260 notes for the significant concepts, and 83.46% were true signals, which means mentioning the concepts in excepted attributes.

Discussion
Our work describes a high-throughput and reproducible approach (ConceptWAS) to use EHR notes to early identify pandemic disease symptoms and investigate clinical manifestations for further hypothesis-driven study. We examined clinical concepts distinguishing COVID-19-positive patients from negative controls and ran the analysis through cumulative time windows. Our results showed that ConceptWAS detected the loss of smell and taste three weeks before CDC added smell and taste disorders to its list of COVID-19 symptoms, demonstrating the power of the EHR to enable early disease detection, and its potential for improving our clinical and public health response to the pandemic.

Code availability. Up-to-date developments of ConceptWAS are available in GitHub (https://github.com/zhaojuanwendy/ConceptWAS).

References
Modeling Variant Annotation (VA) Therapeutic Association Statements using a SEPIO-Based Workflow

Yiqing Zhao, Ph.D.¹, Matthew Brush, Ph.D.², Hongfang Liu, Ph.D.¹, Robert Freimuth, Ph.D.¹*

¹Department of Artificial Intelligence and Informatics, Mayo Clinic, Rochester, MN.
²Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University, Portland, OR.

Background

The Global Alliance for Genomics and Health (GA4GH) is a standards development organization that aims to enable sharing of genomic data and knowledge. The GA4GH Variant Annotation (VA) Project seeks to develop a common data model to represent annotations made about genomic variants, and the evidence and provenance that underlie those annotations. The model will support a wide variety of variant annotations, including causal association to disease/phenotype and interpretations of clinical relevance and actionability, and it will support existing clinical lab standards such as the ACMG/AMP variant interpretation guidelines. As a step toward a generalized VA model, we developed and implemented a detailed process to model one specialized type of VA statement: the Therapeutic Association Statement.

Methods

The modeling procedure adopted the Scientific Evidence and Provenance Information Ontology (SEPIO) workflow(1). The first two steps of SEPIO workflow were: 1. Conduct domain analysis through a comprehensive survey of relevant knowledgebases and clinical databases, collecting use cases in the domain; and 2. Identify data elements from the domain that are required for modeling use case. We examined database structures of existing knowledgebases including PharmGKB, ClinVar, CIViC, OncoKB, MyCancerGenome, PharmVar, and CPIC guidelines. Unstructured statements/assertions curated by those knowledgebases were extracted using natural language processing, which was used to analyze the phrases and data elements in those statements. That analysis identified several statement subtypes, which were modeled in detail.

Results

In our study, we identified five statement subtypes, shown in Table 1.

<table>
<thead>
<tr>
<th>Statement Subtype</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Therapeutic Response (TR) Association</td>
<td>An association between a genetic variation and the response of a patient or other biological system to a therapeutic intervention - specifically, the extent to which an intended benefit is achieved, or an unintended consequence is manifest</td>
</tr>
<tr>
<td>Treatment Course (TC) Association</td>
<td>An association between a genetic variation and the manner in which a treatment is most effectively applied (e.g. the timing, duration and dose of treatment)</td>
</tr>
<tr>
<td>Pharmacokinetic (PK) Association</td>
<td>An association between a genetic variation and how a drug is processed by body (including its uptake, distribution, metabolism, elimination, as well as its concentration in the body)</td>
</tr>
<tr>
<td>Molecular Response (MR) Association</td>
<td>An association between a genetic variation and the expression, activity, or functions of a specific gene product following a treatment</td>
</tr>
<tr>
<td>Clinical Outcome (CO) Association</td>
<td>An association between a genetic variation and a defined clinical outcome that follows a treatment</td>
</tr>
</tbody>
</table>
Based on the five statement types identified, a common data model was developed. The core data model includes four mandatory attributes: Subject (the Variation that the Statement is about), Predicate (the relationship asserted to hold between the Subject and the Object), Object (the Therapeutic Intervention for which the Variation is asserted to have an association, often in the context of a disease), and variantOrigin Qualifier (e.g., germline vs. somatic). Several optional attributes were identified (Table 2) that may link to ontologies, other VA models or shared vocabularies by all VA models.

<table>
<thead>
<tr>
<th>Qualifier</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>treatedCondition</td>
<td>A qualifier that refines the scope of the core statement to apply in the context of a particular genetic condition that is being treated by the intervention.</td>
</tr>
<tr>
<td>population</td>
<td>A qualifier that refines the scope of the core statement to apply in the context of a particular population of individuals.</td>
</tr>
<tr>
<td>comparator</td>
<td>A qualifier that extends the core statement to report a variation or genetic state relative to which the association asserted for the subject Variation holds as true.</td>
</tr>
<tr>
<td>specificSideEffect</td>
<td>A qualifier that extends the core statement to capture a specific unintended consequence associated with the presence of the subject variant, following the indicated treatment.</td>
</tr>
<tr>
<td>clinicalAspect</td>
<td>qualifies a side effect by describing a specific aspect of the side effect that is associated with the subject variation (e.g. its severity, onset, rate of progression, etc.)</td>
</tr>
<tr>
<td>affectedGene</td>
<td>For the MolecularFunction Response Statement type - a qualifier that extends the core statement to capture the specific gene/product being described</td>
</tr>
<tr>
<td>clinicalOutcome</td>
<td>For the Clinical Outcome Association Statement type - a qualifier that extends the core statement to capture a specific clinical outcome associated with the variant (typically an unintended side effect).</td>
</tr>
<tr>
<td>responseMeasure</td>
<td>A qualifier that captures the specific trait/phenotype/characteristic measured to assess response to treatment. This may be a measure to determine if the treatment is having its intended effect (typically a direct symptom or marker related to the condition the drug is given to treat). Or a measure to determine the presence/extent of an unintended side effect.</td>
</tr>
</tbody>
</table>

**Conclusion**
We analyzed curated statements from several community knowledgebases and proposed a VA Therapeutic Association Statement model. The core model follows subject, predicate, and object structure and eight qualifiers were identified in addition to the core model. We are currently expanding the validation of this model using content from clinical genetic reports. The impact of our work is two-fold: 1. the validated model could potentially facilitate integration of genomic medicine information and knowledge into clinical systems with a high-level of consistency, scalability, and computability, and 2. this model will inform the development of a generalized VA model that extends beyond the scope of Therapeutic Association.

**Reference**
Lessons Learned from an Enterprise-Wide Clinical Datathon

Andrew J. Zimolzak, MD, MMSc1,2, Jessica A. Davila, PhD1,2, Vamshi Punugoti, MS1, Katherine H. Sippel, PhD1, Ashok Balasubramanyam, MD,1 Paul Klotman, MD,1 Laura A. Petersen, MD, MPH,1,2 Ryan H. Rochat, MD, PhD1,3 Gloria Liao, MS,1 Rory R. Laubscher, PhD,1 Lee Leiber, MBA,1 Christopher I. Amos, PhD1

1Baylor College of Medicine, Houston, Texas; 2IQuEST and Michael E. DeBakey VA Medical Center, Houston, Texas; 3Texas Children’s Hospital, Houston, Texas

Introduction

“Hackathons” are focused events used to develop or improve software. Recently, this concept has been applied to healthcare data science1 in the form of datathons or data challenges. Baylor College of Medicine (BCM) held a datathon in 2020 to showcase our ability to perform quality improvement and population health studies and launch new research initiatives. Our goals were to (1) inform potential users of available data, (2) leverage local data to address clinical questions, (3) test local data and identify limitations of data use, and (4) facilitate cross-institutional collaborations among faculty, postdoctoral trainees, and students.

Methods

BCM is a leader in medical education, supporting the training of over 3,000 medical and graduate learners annually. Its faculty support numerous healthcare organizations. These include Baylor St. Luke’s Medical Center (BSLMC, a teaching hospital) and a large Faculty Group Practice (FGP) for outpatient visits. These organizations use separate electronic health records and have variable degrees of data governance and data access. The BCM data warehouse, an integrated data repository, is hosted as an Epic Caboodle database. It houses clinical data on 4.3 million unique individuals, 151 million encounters, and 122 million lab tests from February 2009 to the present.

Datathon planning began in Fall 2019. The event was championed by BCM leadership and the leaders of the FGP and BSLMC. It was endorsed and supported by the Chief Information Officer and other informatics leaders. We established a datathon planning committee to determine event policies, facilitate collaborations, and identify potential projects. The planning committee reviewed preliminary data from the BCM data warehouse, and four test case projects were conducted to ensure data delivery processes and usability. Between April and July 2020, we developed approaches for submitting and reviewing proposed projects. We solicited applications to participate in August 2020, with a particular focus on reaching trainees (students, residents, fellows), and a panel of senior faculty from across the College reviewed the applications. Because we conducted the datathon as healthcare operations and identifiable data if needed, only participants within BCM were eligible. They must file a protocol with the institutional review board to continue analyses or present any generalizable knowledge discovered. The committee selected proposals based on feasibility and relevance to quality improvement of population health. Although the COVID-19 pandemic affected the initial plans for an in-person event, we transitioned to a virtual format, and no delays in planning or hosting the datathon occurred.

Datathon projects were vetted for feasibility and impact and selected in September. Data extraction began immediately. Project teams submitted data requests to BCM Information Technology (IT) and received a schedule for data availability. Teams had iterative consultations with IT to refine parameters for data extraction. Microsoft Teams was used for real-time collaboration. Project teams began analysis when data were available. Consequently, some groups had more time for analysis than others. Our datathon was innovative because it involved sustained effort rather than a few days of intense work. Most analysis took place in October, with a preliminary presentation of lessons learned on October 8. IT provided a final dataset to teams for analysis; final presentations were on October 27. A six-person judging committee, including faculty of the affiliated hospitals, gave four awards (most clinically innovative, most innovative use of data, excellence in collaboration, greatest potential for impact on patient care). The judging committee and coauthors determined opportunities and challenges by reviewing participants’ work and comments during the datathon. A survey of datathon participants is also underway to identify areas of success and opportunities for improvement.

Results
Altogether, we received 33 proposals and selected 13 (3 outpatient, 8 inpatient, 2 combined inpatient/outpatient). The following topic areas were covered: early warning for acute kidney injury (AKI) after surgery, seasonality of AKI and respiratory viruses, designing surgical instruments using radiographs, characteristics of COVID patients with complications, referrals and quality in chronic kidney disease using geoanalytics, blood pressure variability and intracerebral hemorrhage outcomes, predictors of COVID outcomes, fluid balance effect on outcomes in patients with subarachnoid hemorrhage, complications of cancer treatment, smoking and cancer screening, inappropriate recording of antimicrobial allergies, and disparities in fragility fracture care between two hospitals.

Discussion

This datathon met our goals by familiarizing investigators with available data and its limitations, familiarizing IT with investigators’ requests, and identifying challenges in the secondary use of clinical data. This will allow BCM to address relevant and timely research questions in the future by utilizing locally integrated data sources.

We uncovered several opportunities through this experience. The greatest opportunities pertained to collaboration. Staff with diverse expertise across the organization engaged in many projects, leading to timely and innovative project ideas. A common platform for collaboration would continue to support interactions between IT and clinicians/researchers. Second, the questions posed by teams were highly dependent on data availability, thus presenting an opportunity for self-service tools like Epic SlicerDicer and i2b2 (only recently broadly deployed at BCM) to provide insight into data availability. Lastly, an executive champion was critical. We hope that this report of a datathon’s benefits will help others decide to use this strategy.

In addition to opportunities, we identified several challenges: many relating to communication within the team itself or the data request and extraction. First, the perception of a data request differs between IT and clinicians. What a provider sees on a daily basis is not reflected in the architecture of the data warehouse. Therefore, outreach and promotion could help manage these expectations. Second, clinicians and IT project members interact and communicate well, but each has expertise that creates barriers to efficient dialogue. This is manifest at times in documentation for some data domains. For example, laboratory studies exist in a database table called “procedures,” a label clinicians would not typically apply. Taking the names of tables and fields at face value runs the risk of overlooking important data or including irrelevant data. This discrepancy in technical language consequently made data request/acquisition and analysis more time-intensive in our projects. Interactive tools or staff with cross-training in clinical and IT domains could improve efficiency. While these challenges were greater with less experienced users, team members with expertise in clinical and IT fields could steer projects efficiently. Third, we found that project tools were a challenge. Tools such as i2b2 and Epic SlicerDicer can retrieve and analyze data, but not all teams were familiar with them. Agreeing upon a common set of tools at the beginning of a project coupled with targeted instruction can improve understanding and support efficient transitions between project tasks. Fourth, most projects required extensive data cleaning, a common phenomenon when using clinical data for secondary analyses. Increased integration of data sources could reduce time and effort spent on data cleaning, but projects still need to budget time for cleaning. Integration or patient linkage across data sources should also be automated as much as possible. Fifth, large, high-resolution datasets (e.g., critical care vital signs) were challenging to curate; however, these granular data were valuable when analyzed appropriately. Also, the demands of image retrieval and analysis suggest a need for a dedicated platform for imaging analysis and facile data transfer in the future.

We have illustrated a new use of a datathon: to “kick-off” a new data warehouse and familiarize users with it. We will continue to learn from challenges and leverage opportunities to help BCM and others use their valuable clinical data.

References

The Role of eHealth Developers and Medical Professionals in Reducing Ageism and Geriatric Digital Exclusion

Dani Zoorob MD MHA MBA MHI, Yasmin Hasbini MD, Victoria Wangia-Anderson PhD MS, Brian Miller MD, Debi Brobst MBA

Introduction
Even though the field of telemedicine and e-health is expanding, the rate of uptake of telemedicine among the elderly does not match the expansion rate, leading to disparities and inequity in care availability between populations. This phenomenon contributes to the digital inverse care law, whereby those who need the services the most (in this case, the elderly with the service in question being digital medicine) are in fact excluded from receiving it, and thus result in worse outcomes.1 This results in digital exclusion and ultimately ageism.

Ageism is expressed in two ways: the lack of accommodations specific to the elderly during the development of telemedicine platforms, and in providers presuming lack of interest or ability to use digital technology by the elderly.2,3 When experienced, perceived ageism leads to lower uptake of digital medicine by the older population.4,5 This leads to digital disparity and exclusion, which is now being characterized as a social determinant of health due to it being a social and economic factor impacting health outcomes.1 Thus, the objective of this study was to identify the motivators, concerns, and issues contributing to the uptake (or lack of) of e-Health platforms among the elderly in order to propose clear and quantifiable measures that developers and providers can consider to decrease access inequity.

Methods
This was a cross-sectional study whereby an anonymous survey was administered to patients at a Urogynecology and Menopause Center. A total of 205 completed the survey (225 total surveys, 93% response rate). The survey was divided into sections: motivators of use of e-health platforms, concerns and issues experienced when using them, enhancements desired in the platform, and women’s health specific questions for insight into care planning.

Results
The average age of the participants was 68.9 (SD 4.9). The surveyed population was mostly Caucasian (83.9%), with 11% Hispanics/Latinas and 3% African Americans, while up to 80% had a college degree. Privacy, cost, and technical issues were evaluated (Fig 1. a) with the majority being concerned about running into technical issues (50%) and privacy (45%). The most encountered technical issue was that the system requested much information upon startup (36%), followed by being unable to print/save/send documents from EHR (35%) (Fig 1. b). Of those above the age of 75 years, 25% experienced difficulty from physical limitations such as arthritis and finger mobility, and 20% complained of eye strain/joint. Experiencing anxiety when using portals was reported in up to 36% of the participants (Fig 1. c). The telehealth portals feature changes requested included the ability to freely change the size of the icons and text (47%), a simpler design (46%), and more drop-down lists to choose from (40%) (Fig 2. a).

Up to 40% of those surveyed reported that portals make their medical visits less stressful, and 24% specifically correlated that to logistics. Regarding condition-specific telehealth visits (Fig 2. b), 74% were comfortable with medication checks via telehealth, 66% were interested in post-operative visits being online (excluding those requiring a pelvic exam), and 52% verbalized preference for preoperative counseling via telehealth. When questioned about the potential motivators for using technology-enabled care, 73% of the participants reported that they would consider portal use if specifically encouraged by their providers (Figure 2. c)

Discussion
As the population ages, the proportion using the portals will be increasing significantly within the next two decades. Ensuring uninterrupted healthcare access to these patients is necessary - although this is gradually being addressed since the pandemic started. The study suggests that the health informatics field has not yet reached adequate accessibility maturity, specifically from an ageism standpoint. Although platforms are becoming more user-friendly, their tailoring of age-focused accessibility options has been limited. Our study suggests that when designing for such a population, it is essential to incorporate a flexible and modifiable platform that can be changed to the patient’s desires or at least adapt to the age group. Although only the older segment of our population complained of physical or medical limitations impacting their use of portals, more complained of anxiety when using technology, indicating that more time and effort should be spent educating and assisting patients in understanding how to use the platforms and addressing concerns.
Additionally, the promotion of portal use may be hindered by the ageism inherent in the behavior of providers and medical staff. An important outcome of our study is the indisputable role providers - and healthcare systems as a whole - could play in increasing uptake of e-Health among the geriatric population. This is through direct promotion by providers, active patient education, and on-site enrollment. Additionally, there is a clear consensus on the importance of addressing privacy concerns when developing the educational material. Since our study was administered to urogynecology patients, it is essential to note the versatility of e-Health applicability, as indicated by our results of a generally positive attitude. This can be expanded to other medical and surgical specialties.

**Conclusion**

Tackling ageism in health informatics is necessary. Accessibility and adaptability in portals and platforms must include age-related needs. Policymakers and healthcare leaders should develop a more standardized method of educating and enrolling their geriatric population in e-health services, as well as investing in developing more flexible interfaces that can be modified to the patient’s desire.

**References**


Network Analyses on Hyperlinked Online Resources for COVID-19 Vaccine

Meredith Abrams, MS, MPH; Dongwen Wang, PhD
Biomedical Informatics, Arizona State University, Phoenix, AZ

Introduction
Since the outbreak of the COVID-19 pandemic, the related online information resources have been exploding and constantly changing. Timely collection and analysis of these resources is critical for effective dissemination of the latest clinical evidence. Here we report a pilot study to develop a process to analyze the hyperlinked online resource networks for COVID-19 vaccine and to assess the change of the networks over time. This work is part of a methodology development to integrate information “freshness” into network analysis for timely dissemination of clinical evidence.

Methods
We developed a webcrawler to automatically collect the hyperlinked online resources based on a list of 252 keywords related to various topics of COVID-19. On Feb. 15, 2021, we started the first webcrawling from the COVID-19 resource pages of the CDC and WHO websites (the top authorities of related information), recursively searched the hyperlinked resources to a depth level of 4, and constructed networks at the resource and the hosting site levels. To narrow down the resources to focus on COVID-19 vaccine, we processed the network data, only kept resources matching with keywords related to COVID-19 vaccine, and constructed a subnetwork. We then visualized the networks and assessed their metrics using the Gephi tool. We examined the top resources, the associated keywords, and the hosting sites for both the overall network and the subnet. We repeated this process on Mar. 4, 2021, and assessed the changes of the networks over time.

Results

<table>
<thead>
<tr>
<th>Overall Network of COVID-19</th>
<th>Subnetwork of COVID-19 Vaccine</th>
</tr>
</thead>
<tbody>
<tr>
<td>unique resources</td>
<td>hosting sites</td>
</tr>
<tr>
<td>Feb 2021</td>
<td>158,778</td>
</tr>
<tr>
<td>Mar 2021</td>
<td>610,990</td>
</tr>
</tbody>
</table>

For the COVID-19 overall network, we found the top keyword hits across February and March included CDC, COVID-19, vaccine, R0, and treatment. The top individual resources based on PageRank were sponsored by government, such as the homepage of USA.gov, the homepage of HHS.gov, and the CDC’s YouTube channel. The top hosting sites included government agencies, such as CDC, and social media platforms, such as Facebook, Twitter, and Instagram. The COVID-19 vaccine subnetworks are visualized in Figure 1.

With regard to the change of the overall network, we recorded significant increase of both the unique resources (284%) and hosting sites (215%) between February and March. Only 33% (51,785 out of 158,778) of the unique resources and 40% (2,518 out of 6,374) of the hosting sites identified in February remained in the network in March. Meanwhile, 92% (559,205 out of 610,990) of the unique resources and 87% (17,562 out of 20,080) of the hosting sites in March were newly identified. For the COVID-19 vaccine subnetwork, 72% (809 out of 1,126) of the unique resources and all (100%) the hosting sites identified in February remained in the subnetwork, and 40% (547 out of 1,356) of the unique resources and 20% (21 out of 104) of the hosting sites in March were newly identified.

Discussion and Conclusion
We successfully explored a process to construct an overall network of hyperlinked online resources for a specific domain (COVID-19) and a subnetwork for a specific topic (COVID-19 vaccine). We identified the most popular resources and hosting sites, and categorized the resources based on specific keywords. We observed a radical change of the overall network for COVID-19 over a period of three weeks, but a more stable subnetwork for COVID-19 vaccine. To account for the dynamics of the networks over time, we are now developing measures on information “freshness” and incorporating them into online resource network analysis for timely dissemination of the latest clinical evidence.
Exploring the impact of perceived usability and cognitive workload on objective performance using supervised machine learning classifiers

Karthik Adapa, MBBS, MPH\textsuperscript{1}, Shiva Das, PhD\textsuperscript{1}, Prithima Mosaly, PhD\textsuperscript{2}, Fei Yu, PhD\textsuperscript{2}, Carlton Moore, MD\textsuperscript{1}, Lukasz Mazur, PhD\textsuperscript{1,2}

\textsuperscript{1}School of Medicine, \textsuperscript{2}School of Information and Library Science, UNC-Chapel Hill, North Carolina, USA

Introduction

In health informatics, suboptimal usability of electronic health records (EHRs) and cognitive workload (CWL) of healthcare professionals (HCPs) are known to independently impact performance (e.g., increased likelihood of medical errors)\textsuperscript{(1)}. However, limited studies have examined the combined effect of perceived usability and CWL on the objective performance of HCPs\textsuperscript{(2)}. In radiation oncology, most medical centers typically develop and implement Quality Assurance (QA) checklists to reduce errors but often do not evaluate the impact of usability and workload on HCP’s performance\textsuperscript{(3)}. Therefore, we explore the combined impact of perceived usability and CWL of a quality assurance (QA) checklist on dosimetrists’ performance during the radiation therapy planning process for cancer patients.

Methods

Dosimetrists performed QA tasks on 73 different types of radiation therapy plans with varying degrees of difficulty (low (35), medium (35), and high (6) based on subjective rating by physicists). The perceived usability of the QA checklist was assessed using the System Usability Scale (SUS) and Post-study System Usability Questionnaire (PSSUQ). The perceived CWL was assessed using National Aeronautical Space Agency Task Load Index (NASA-TLX) and Workload Profile (WP). Performance was measured objectively as the time taken to complete QA tasks (with two classes: low-22 and high-51 based on the average time taken to complete the checklist). To correct the class imbalance, we used the synthetic minority oversampling technique and randomly sampled the minority to achieve the same size of majority class (low-51 and high-51). The usability (SUS & PSSUQ) and workload (NASA-TLX & WP) measures were treated as independent variables and employed individually and in combination to create models aimed at predicting the two classes of objective performance. Five different supervised machine learning algorithms that work best for binary classifications were adopted to predict performance a) Naive Bayes(nb) b) k-Nearest Neighbors(knn) (k=5) c) Logistic Regression (lr) d) Support Vector Machine (svm) e) Random Forest (rf). We used 10-fold stratified cross-validation (CV) strategy and kept the same training and test sets across the classification techniques and the different combinations of independent features to ensure a fair comparison of the different trained models using the same data of training and test sets.

Results

Most of the combined variable models (highlighted in blue; see Table 1) achieved higher average accuracy than the single variable models. Except for knn, among single variable models, SUS or WP were worse in predicting objective performance across the other four classifiers. However, the addition of NASA-TLX and PSSUQ respectively significantly enhanced the prediction of objective performance.

Conclusion

Preliminary findings from this on-going study suggest that centers implementing QA checklists may employ usability and CWL jointly to predict performance. Future research will build on these findings by expanding the sample size, utilizing multiple validation strategies, testing for statistical significance, among others.

References

Assessing Data Quality of Higher Frequency Time Series Vital Signs Captured in Intensive Care Units

Ali S. Afshar, PhD¹, Yijun Li, MSc², Zixu Chen, B.Sc.³, Yuxuan Chen, B.Sc.³, Jae Hun Lee, B.Sc.³, Darius Irani, B.Sc.³, Aidan Crank, MSE³, Digvijay Singh, MSE³, Michael Kanter, MD⁴, Nauder Faraday, MD, MPH¹, Hadi Kharrazi, MD, PhD⁵,⁶

¹Johns Hopkins School of Medicine, Department of Anesthesiology and Critical Care Medicine, Baltimore, USA
²John Hopkins School of Public Health, Department of Epidemiology, Baltimore, USA
³Johns Hopkins Whiting School of Engineering, Department of Computer Science, Baltimore, USA
⁴Kaiser Permanente Bernard J. Tyson School of Medicine, Department of Clinical Science, Pasadena, USA
⁵John Hopkins School of Public Health, Department of Health Policy and Management, Baltimore, USA
⁶Johns Hopkins School of Medicine, Division of Health Sciences Informatics, Baltimore, USA

Introduction

Physiological signals represent a major source of information in intensive care units (ICUs). Previous studies have addressed data quality issues for vital signs routinely entered in EHRs by healthcare practitioners. However, those studies have not assessed data quality for large-scale capture of higher frequency physiologic signals that is now possible within ICU settings. In this study, we explore the data quality issues, specifically completeness, accuracy and timeliness, of higher frequency (minute-by-minute) time-series vital signs data captured in ICU settings.

Methods

We used Medical Information Mart for Intensive Care (MIMIC-III) database¹. We measured the completeness, accuracy, and timeliness of four time-series (minute-by-minute) vital signs data streams: heart rate (HR), respiratory rate (RR), blood oxygen saturation (SpO2), and arterial blood pressure (ABP), in MIMIC-III. Completeness of time series vital signs data was measured: (1) by the presence of a non-NULL record for a given physiologic variable (HR, RR, SpO2, ABP) corresponding to an ICU stay for a patient, and (2) by the proportion of non-missing data for each physiologic variable in a given ICU stay for a patient. If an entire record corresponding to a physiologic variable was non-existent or NULL, then the corresponding physiologic variable for that ICU stay for the patient was flagged as incomplete. Accuracy of the vital signs was measured by a series of rules examining the physiological limits of the vital signs. Timeliness of the vital signs was assessed as the overlap of the minute-by-minute records with the ICU stay.

Results

Of the 22,247 time series vital sign records in the MIMIC-III Waveform Database Matched Subset¹, 21,326 records were: (a) mappable to and representative of distinct ICU stays; and, (b) associated with patients aged 16 or older. The 21,326 vital signs records corresponded to 16,009 distinct ICU stays and 9,410 unique patients. We applied certain filtering steps associated with timeliness, completeness, and accuracy of vital signs with a minimum duration of one minute being collected during the ICU stays. When these filtering steps were applied, the minimum one-minute HR, RR, and SpO2 records were found for most patients (>98%), but only 7,496 (49.6%) of patient-ICU stays had ABP records available. We observed a ~28% drop in the number of HR, RR, and SpO2 records per ICU stay after applying the timeliness constraint (i.e., having at least one minute overlap with the vital sign’s associated ICU stay). This percentage for ABP time series records was ~13%. We finally explored the proportion of ICU stays (0 to 1) against the percentage of ICU length of stays (0% to 100%) that were covered by acceptable vital signs data (i.e., complete, accurate, and timely) in the study population (Figure 1). The proportion of patient-ICU stays that had 100% of their length covered by an acceptable vital sign data were 9.2% for HR, 8.8% for RR, 7.9% for SpO2, and 3.9% for ABP (represented by the lowest point of each vital sign line in Figure 1).

References

Detecting Major Depressive Disorder from Clinical Notes using Neural Language Models with Distant Supervision

Bhavani Singh Agnikula Kshatriya, MS1, Nicolas A Nunez, MD3, Manuel Gardea-Resendez, MD3, Euijung Ryu, PhD2, Brandon J Coombes, PhD2, Sunyang Fu, MPH1, Mark A Frye, MD3, Joanna M Biernacka, PhD2, Yanshan Wang, PhD1,4
1Department of Artificial Intelligence & Informatics, Mayo Clinic, Rochester, MN; 2Department of Quantitative Health Sciences, Mayo Clinic, Rochester, MN; 3Department of Psychiatry & Psychology, Mayo Clinic, Rochester, MN; 4Department of Health Information Management, University of Pittsburgh, Pittsburgh, PA

Introduction: Major depressive disorder (MDD) is a prevalent psychiatric disorder and is associated with significant healthcare burden worldwide1. Although structured Electronic Health Records (EHRs), such as ICD-9 or ICD-10 codes, have been used to represent MDD phenotypes2, lots of MDD phenotypic information is documented in free-text EHRs, such as clinical notes. Using natural language processing (NLP) techniques to detect MDD phenotypes from clinical notes are vital to phenotype patients with MDD. In the literature, only a few studies utilized clinical notes for phenotyping MDD, and, to the best of our knowledge, no studies deployed deep learning based neural language models for detecting MDD from clinical notes. In this study, we applied the state-of-the-art neural language models, namely BERT and Bio-Clinical BERT within a distant supervision framework to identify MDD using clinical notes and compared with conventional machine learning models.

Method: In this study, we used the Mayo Clinic Biobank cohort to build a corpus of clinical note documents. We first used MDD-related ICD codes to identify cases and controls from the biobank cohort. We randomly selected 500 cases and 500 controls, and retrieved clinical notes from the Mayo Clinic Data Warehouse. This resulted in 168,139 clinical note documents. The goal is sentence-level identification of MDD instances, i.e., Positive (e.g., “Patient has history of dysthymia”), Possible (e.g., “Patient is a depression suspect”), Negative (e.g., “There is no evidence of depression”), or unknown MDD instances, for each sentence in each clinical note document. We randomly selected 500 documents as a gold standard testing dataset and asked two psychiatrists to manually annotate MDD at the sentence level. For the remaining 167,639 clinical note documents, we developed a rule-based NLP algorithm to generate weak labels. This weakly labeled data was used as training data to train neural language models and machine learning models. Since the dataset was highly imbalanced with 98% unknown and 2% (i.e., 53,569) MDD related instances out of total 3,081,359 sentences, we performed under-sampling to balance the data and created a final training dataset of 107,138 sentences. We have used pretrained BERT base model and Bio-Clinical BERT with an additional dropout along with a classifier layer with output dimension of four. Cross entropy loss function was used with the Adam optimizer. For the conventional machine learning models, we tested Random Forest (with n=100 estimators), K-NN (with K=7), SVM (with linear kernel) classifiers with word embeddings as the input features. We used Precision, Recall and F1 score metrics to evaluate the models.

Results: Bio-Clinical BERT has achieved the best performance for all defined MDD classes with F1 scores of 0.99 for unknown, 0.95 for positive MDD, 0.87 for possible MDD, and 0.91 for negated MDD. BERT is inferior to Bio-Clinical BERT but still performed better than machine learning models.

Conclusion: The results indicate that Bio-clinical BERT with domain specific clinical contextual representations performs better for clinical NLP tasks. Since the models were evaluated on manually annotated gold standard, it shows that the distant supervision is an effective technique to be combined with neural language models for detecting MDD from clinical notes which consequently can reduce a lot of efforts and time for manual annotation.

References
Building a Self-Service Analytics Ecosystem for Program Data Sharing and Dashboards (NIH ‘All of Us’ Precision Medicine Initiative)

Authors: Toufeeq Ahmed PhD¹, Robert Abram BS¹, Bhinnata Piya MPH¹, Keri Wolfe MS¹, Dylan Klomparens MS², and Paul Harris PhD¹.  
1 Vanderbilt University Medical Center, Nashville, TN  
2 National Institutes of Health, Rockville, MD

Summary: The All of Us Research Program (AoURP) has embarked on a historic mission to build a comprehensive dataset that includes health information shared by a diverse cohort of at least one million participants across the United States. In addition to making data available for approved researchers to advance precision medicine, the AoURP will also make research results accessible to participants. With the growing need to understand program operations and related monitor key performance indicators (KPIs) for strategic decision-making, the Vanderbilt University Medical Center (VUMC) Data and Research Center (DRC) has worked with the NIH and other AOURP program partners to create a self service analytic dashboard ecosystem and business intelligence platform for review of program operations data using Tableau server on the frontend and with Apache NiFi and PostgreSQL in the backend. This scalable, access and privacy enabled design is created to facilitate our platform as a “Bring Your Own Data (BYOD)” model, enabling our technology partners to bring in operational data using open source ETL tool Apache NiFi into dedicated PostgreSQL (open-source) datamarts to ingest, transform, and create custom analytics dashboards using Tableau server.

The DRC Analytics Ecosystem includes dashboards to display program analytics support program awareness and decision making, operational analytics to support business operations, and partner operations dashboards which are created by AoU consortium partners to support their business operation. Tableau Server and Google BigQuery were selected for their best-in-class offering for BI/data visualization and cloud-based data storage, respectively.

This poster presentation will illustrate 1) a model that research institutions can adopt to build a robust and collaborative analytics ecosystem, 2) the strengths of free, open-source systems such as Apache NiFi and PostgreSQL, and 3) the challenges associated with implementing an analytics platform in a FISMA-moderate cloud infrastructure.

Conclusion: Open source tools, such as Apache NiFi and PostgreSQL, can play a key role in enabling research institutions of all sizes to adopt a data sharing model that is scalable, robust, and collaborative.
Machine Learning for Predicting Hospital Acquired Pressure Injuries in ICU Patients: From Explainable AI to Ensemble Super Learners

Jenny Alderden¹ Ph.D. APRN; Andy Wilson¹² Ph.D. MSTAT; Sergey Krikov¹ MS; Ryan Butcher¹ MBA; Jonathan Dimas¹ Ph.D. RN; Tracey Yap³ Ph.D. RN FGSA FAAN
University of Utah¹ Salt Lake City UT; Parexel² Durham NC; Duke University³ Durham, NC

Background: Hospital-acquired pressure injuries (HAPrIs) occur on the skin caused by pressure in 6-10% of ICU patients, and annual costs for treatment exceed $65 billion annually. Preventing HAPrI requires accurate risk assessment to enable nurses to allocate preventive resources. Currently available tools such as the Braden Scale tend to classify nearly all ICU patients as ‘high risk’ and therefore have limited clinical utility in the ICU. Advances in machine learning (ML) present an opportunity to modernize and improve HAPrI risk assessment using Electronic Health Record (EHR) data. Yet, it is not possible to know a priori which ML algorithm will perform best. Because there is no single best-practice technique, we used a super learner approach to stack multiple ML models into a composite model with enhanced predictive accuracy. Explainable AI was also leveraged to understand how specific features in the model influenced the predicted outcome.

Methods: For 5,101 University of Utah surgical and cardiovascular ICUs patients, including 399 (7.8%) with HAPrI, we applied machine learning best practices, including addressing informative missingness in data, class imbalance, and leveraging multiple models with ensemble (super learner) models. Inclusion criteria were age ≥18 and admission date between 2014-2018. All analyses were performed in open source software (R v 4.0.3 via H2o v 3.32.0.1 packages). The code will be made available via a dedicated open science framework (OSF) page for reproducibility and transparency.

Results: Ensemble (super learner) models outperformed all individual models (AUC 0.862 for the best base learner [xgb] vs. 0.878 for ensemble) on a 20% held-out test set. All models modestly improved after minority oversampling. Partial dependence plots produced for each variable indicate differences in individual models’ use of predictor variables (Fig 1).

Conclusions: Ensemble ML methods such as super learners are a useful tool for predicting HAPrIs, particularly when compared to current tools, which tend to identify all patients as ‘high risk.’ The next steps include external validation and subsequent implementation with associated clinical decision support.
Colonizing Microbiome as a Determinant of COVID-19 Outcome: A Pilot Study

Alexander V. Alekseyenko, PhD, FAMIA, Bashir Hamidi, MPH, Stephane Meystre, MD, PhD, FACMI, FIAHSI, FAMIA
Medical University of South Carolina, Charleston, SC, USA

COVID-19 is characterized by immune modulation following exposure to the SARS-CoV-2 virus. This condition in the worst cases results in polymicrobial infectious complications involving bacteria and possibly fungi. Surprisingly, not all patients follow the same course, and many have little symptoms and/or no complications. We hypothesize that the colonizing microbiome is linked to the clinical status of the COVID-19 patients and may provide early leads for stratification of cases into risk categories. The rationale for this hypothesis comes from the following facts: (i) colonizing opportunistic pathogens may turn infectious following immunosuppression; (ii) diversity of the colonizing microbiome has been shown to correlate with disease severity in many conditions; (iii) microbiome diversity can be viewed as a robust proxy of the host immunological state, with loss of diversity indicating a severe infectious state and proneness to medical complications. The conceptual model for the premise of this study is shown in Figure 1.

At MUSC Program for Human Microbiome Research (PHMR), we are in a strong position to evaluate our hypothesis. First, we have established IRB- and IBC-approved protocols for access to surplus COVID-19 testing materials using the Living uBiome Bank (1). Currently, we have biobanked 1,160 positive and negative specimens from these materials and we anticipate prospective access to over 400 additional specimens per day. Second, we have demonstrated that useful microbiome information is available in similar surplus specimens. We have previously obtained preliminary data on the composition of bacteria co-colonizing with methicillin resistant Staphylococcus aureus and vancomycin resistant Enterococci. These data have been obtained from surplus infection surveillance swabs, similar to those used for COVID-19 testing. Finally, our current MUSC IRB grants us access to medical records to allow for precise interpretation the microbiome compositions, which is a major plus compared to many existing microbiome research studies, where availability of such detailed clinical information is often a major limitation. Access to detailed clinical information is further enhanced by the recently developed COVID-19 clinical data mart under the direction of the Translational Biomedical Informatics Center (TBIC).

Beyond immediate clinical significance of our project, organization of the relevant clinical data using the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) standard will allow for future linkage into national and international research efforts. This project will generate new data on composition of commensal microbiota and how it relates to susceptibility to SARS-CoV-2 infection and complications in COVID-19 disease course; little to nothing is currently known about this. Integration of the EHR clinical data with high-throughput molecular data (like the microbiome) is an urgently needed and actively developing area of quantitative methodological research. Use of the OMOP CDM with clinical data as well as microbiome molecular data is innovative and would contribute to preliminary data on a novel application of the G-CDM extension for genomic information.

References

Evaluating the Utility of a Prototype Clinical Decision Support Tool for Chronic Pain Treatment Choices in Primary Care

Katie S. Allen, BS1,2, Elizabeth C. Danielson, PhD3, Sarah M. Downs, MPH1, Olena Mazurenko, MD, PhD1, Julie Diiulio, MS4, Burke W. Mamlín, MD1,5, Christopher A. Harle, PhD6

1Fairbanks School of Public Health, Indiana University, Indianapolis, IN; 2Regenstrief Institute, Inc., Indianapolis, IN; 3Northwestern University, Evanston, IL; 4Applied Decision Science, LLC, Cincinnati, OH; 5Indiana University School of Medicine, Indianapolis, IN; 6University of Florida, Gainesville, FL

Introduction

Patients suffering from chronic conditions often require clinicians to complete complex tasks related to identifying and choosing the most promising treatments1,2. To address the management of chronic pain, we designed a prototype decision support tool called the Chronic Pain Treatment Tracker (Tx Tracker)3. The Tx Tracker is intended to be a point-of-care tool that summarizes an individual patient’s treatment history and new treatment options based on history, trends in health outcomes, and treatment-related risks. The objectives of this study were to (1) evaluate the perceived utility of Tx Tracker for meeting clinicians’ needs; (2) evaluate the perceived utility of Tx Tracker for identifying pain treatment options; and (3) identify clinician preferences for layout and visual design of the Tx Tracker.

Methods

We conducted a qualitative system design and evaluation study with twelve primary care physicians from two health systems in the state of Indiana between February and May 2019. The twelve interviews consisted of background questions, prototype exploration, a use case scenario, and evaluation questions. Two researchers (ED, SD) jointly reviewed their interview notes and created a single, detailed composite note for each section of the user testing interview for each participant. The same two researchers then created a codebook, and then qualitatively coded the text of each note using the codebook. After coding, two researchers (CH, KA) used a process of upward abstraction, to identify emergent themes within and across the coded data that related to each research objective.

Key Findings

The primary theme was clinicians’ preference to use the Tx Tracker for the right patient at the right time point of time in workflow. Clinicians expressed a preference for access to comprehensive patient history at the point of care. Currently, much of this needed history information is found in clinical notes, which limits searchability and is often spread across time, discipline, and locations. While many clinicians indicated the Tx Tracker could help overcome these limitations, some expressed skepticism about having enough time to utilize Tx Tracker. In terms of interface design and layout, one theme was the value of presenting contextual information that links past treatments to other relevant clinical information, such as diagnoses, goals, and patient preferences. Similarly, our analysis indicated the potential value of visually co-locating information on historical treatments with treatment options and supporting interactivity allowing users to. Finally, clinicians expressed a general preference for visual cues, such as color and prominent screen placement of needed information.

Conclusion

The widespread use of EHRs and proliferation of electronic patient information necessitates research on how to best communicate with users at the point of care, especially for patients with complex chronic conditions. Decision support tools, such as the Tx Tracker, may reduce cognitive burden and increase use of lower risk and more effective treatment options. In this study, we found support for tools like the Tx Tracker to meet clinicians’ information needs, aid identification of treatment options, and improve patient-centered care.

References


Extracting Social Variables from Clinical Documentation to Better Facilitate Response to Patient Need

Katie S. Allen, BS\textsuperscript{1,2}; Dan Hood, MPH\textsuperscript{1}; Jonathan Cummins, BS\textsuperscript{1}; Suranga Kasthuriratne\textsuperscript{1,3}, PhD; Peter J. Embi, MD, MS\textsuperscript{1,3}; Joshua R. Vest, PhD, MPH\textsuperscript{1,2}

\textsuperscript{1}Regenstrief Institute, Inc., Indianapolis, IN; \textsuperscript{2}Richard M. Fairbanks School of Public Health, Indianapolis, IN; \textsuperscript{3}Indiana University School of Medicine, Indianapolis, IN

Introduction

The Covid-19 pandemic has underscored the differential impact of social needs on the health outcomes, specifically across racial groups.(1) Much of the analysis examining these disparities utilize race as the variable of interest, however, race as a social construct, does not embody the underlying root causes of many health disparities, including the social disparities created by systematic racism(1). To facilitate closing the gaps on health disparities, we need to consider the underlying social needs driving them. The goal of this project is to develop methodologies for identifying and extracting social needs from clinical documentation.

Methods

Data and Tools. This study leverages the clinical repository of notes from the Indiana Network for Patient Care, which includes approximately 80 hospitals within the State. To develop the algorithms, the team utilized a systematic, iterative process with keyword identification and state machine development. A training set was identified by selecting twelve months of clinical notes from one health system and a testing set was created utilizing six months of clinical notes from a second health system. To create the natural language processing state machines necessary to identify the presence of a social need, the team leveraged Regenstrief Institute’s nDepth\textsuperscript{™} platform.

Keyword Identification. Utilizing one month of the notes, nDepth\textsuperscript{™} produced a list of continuous sequence words (n-grams) for review. This list was filtered utilizing Term Frequency-Inverse Document Frequency measurements, to identify and remove words not contributing value. Each set of n-grams was reviewed to determine whether it connects the topics of interest. Additional keywords were gathered from validated and literature.

State Machine Development. nDepth\textsuperscript{™} indexes and searches the collection of clinical notes based upon the dictionary created by the selected keywords. State machines were created for each of the social needs and were deployed on the training set. An initial manual review of 200 randomly selected notes was completed. Performance was measured by examining positive and negative predictive value (PPV, NPV). Adjustments were considered and, with the state machine dictionaries then updated, another round of reviews completed. The final iteration was against the test set of notes. For each of the top ten note types, a random sample of 50% of machine positives and an equal number of machine negatives were manually reviewed (KA, DH).

Results

For housing instability, the final dictionary contains 54 keywords and includes rules for negation of certain keywords as well as rules for dictionary terms within a certain word count of other keywords (e.g., sleep within 5 words of car). Two iterations were needed to arrive at the final state machine (See Table 1.) For financial insecurity, the final dictionary contains 51 keywords and includes a grammatical rule denoting when a patient is able “to afford” vs. unable “to afford.” Two iterations against the training set were undertaken (See Table 1.)

<table>
<thead>
<tr>
<th></th>
<th>Housing Instability</th>
<th>Financial Insecurity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training Set Run 1</td>
<td>PPV = 71%, NPV = 100% (n=200)</td>
<td>PPV = 83%, NPV = 100% (n=200)</td>
</tr>
<tr>
<td>Training Set Run 2</td>
<td>PPV = 89%, NPV = 100% (n=200)</td>
<td>PPV = 89%, NPV = 99% (n=1025)</td>
</tr>
<tr>
<td>Test Set Run</td>
<td>PPV = 94%, NPV = 99% (n=998)</td>
<td>Pending</td>
</tr>
</tbody>
</table>

Table 1. PPV and NPV results for each state machine iteration

Conclusion

This project demonstrates the feasibility of utilizing natural language processing to identify and extract social needs from clinical text. This is a critical first step in advancing efforts to address the underlying social needs of patients and to address the health disparities created by systematic racism within our existing healthcare practices and tools.

References

Policy Alienation as a Source of HITECH Act-Related Burnout
Joseline Amado,1 Yousef Latif,2 Mustafa I. Hussain, PhD3
School of Public Health, University of California Irvine (UCI); 2School of Business, UCI; 3DMICE, OHSU, Portland

Introduction
Burnout has been recognized as a crisis in US medicine associated with the introduction of electronic health records (EHRs), mandated by the HITECH Act. The EHR has been recognized as a tool which facilitates the limitation of physician autonomy,[1] and in this work we offer policy alienation,[2] previously used in the Netherlands to understand the effects of neoliberal policy reforms on public professionals, as a framework for guiding research on recent US healthcare reform. Here we report early results from a review.

Background
Burnout tends to emerge when one sees a large number of patients while working under political regimens that associated each client or patient with a number of repetitive tasks.[2] Burnout can lead to an increase in resignations and sick leave in addition to creating communication barriers to effective care.[2] See Figure 1 for an overview.

Figure 1. Conceptual overview of policy alienation.[2]

Methods
A structured keyboard search was conducted on several electronic databases, including Cochrane, CINAHL, PubMed Central, and PubMed. Studies on physician burnout in hospitals from 2015–2019, the period following the HITECH incentive sunset, were considered. We identified 37 studies for inclusion in this iteration of the review.

Preliminary Results
The disconnect in policy awareness among physicians stems from their psychological disengagement from the health policy program. There is widespread resentment about the present structure of the EHR and use of CDS that pressures physicians to meet new policies and accreditation guidelines. The pressure to perform the tasks of documentation and care limits physicians’ creative range and discretion with patients, resulting in burnout which may contribute to systemic dysfunction.

Conclusion
EHRs are promising for enhancing care coordination, error reduction, and care quality enhancement. However, these tools in the context of a litigious and metric-driven policy environment has led to consequences such as increased clinical burden, policy alienation, and physician burnout, in the Netherlands and in the United States. Perhaps it is time for technologically-facilitated innovation in the social mechanisms of healthcare governance.

References
Understanding the use of pharmacological knowledge bases in clinical care

Shilo Anders, PhD¹, Laurie L. Novak, PhD, MHSA¹, Nawshin Kutub, PhD², Carrie Reale, MSN, RN-BC¹, Daniel France, PhD, MPH¹, Christopher L. Simpson, MA¹, Courtney Van Houten, MA², Karlis Draulis², Rubina Rizvi, MD, PhD², Tiffani J. Bright PhD², Gretchen P. Jackson, MD, PhD², Anita M. Preininger, PhD²

¹Vanderbilt University Medical Center, Nashville, TN USA; ²IBM Watson Health, Cambridge, MA, USA

Background: Pharmacological knowledge bases (PKB) provide current medication-related information that clinical staff use on a regular basis. They cover a wide variety of drug interactions¹ and are often used to answer pharmaceutical or pharmacological questions including adverse drug reactions, intravenous (IV) compatibility, and other drug-related topics.² Little research has been done on such tools and factors that contribute to clinical use.

Objective: The objectives of this study were to identify medication-related information needs of pharmacological knowledge base users and to examine roles and modes of access of the users, to inform system improvements and feature design. We first conducted a survey of informational needs and usefulness in users’ everyday clinical work. We then interviewed users from the survey to identify specific informational needs sought by users, to gain in-depth insights about how medication-related knowledge bases help them in their day-to-day work.

Methods: Clinical staff that used a pharmacological knowledge base were asked to participate in an electronic survey about information needs and use cases. Subsequently, selected survey respondents were interviewed about their system use, including a walkthrough of their typical practices. Across user types, practice areas, and years of experience, we compared the types and frequencies of information sought, modes of use, and overall experience.

Results: 155 medical personnel completed surveys including nurses (31%), residents (24.5%), fellows (19.4%), pharmacists (8.4%), nurse practitioners (7.1%), attending physicians (4.5%) and others (5.2%). In surveys, users cited accessibility as the most important feature driving use, followed by reliability and general ease of use. We found statistically significant differences (Pearson Chi-square: 20.777, P = 0.008) across user roles and frequency of use of Micromedex; nurses and pharmacists used Micromedex daily or several times per week; nurse practitioners and residents, several times per month; and fellows, less than once per month. There were no significant differences in ratings of Micromedex by role, years in role, or practice area. Figure 1 shows the results what tasks each role used Micromedex for from both the surveys and interviews (N=13). Most interviewees use Micromedex on a larger form factor and resort to smaller tablets and smartphones when moving around, such as during rounds.

Conclusion: Pharmacological knowledge bases are commonly used tools to help hospital staff in medication management. Nurses and pharmacists are the most frequent users of such resources; physicians and nurse practitioners use them much less frequently. Areas of IV compatibility and drug/dosing information are the most common information types sought using a pharmacological knowledgebase to ensure and maintain patient safety. Information needs associated with various user roles and contexts of use can be used to inform future developments and product design that best aligns with user needs and workflow processes.

References
COVID-19 Trial eSource Data Capture Efficiencies using OneSource

Adam Asare, PhD1,2, Alejandra Jauregui1, Andrew Robinson1, Mitra Rocca, PhD3, Cal Collins4, Heidi Collins, MS1, Laura Esserman, MD, MBA1 and Kathleen Liu, MD1

1University of California, San Francisco, CA, 2Quantum Leap Healthcare Collaborative, San Francisco, CA, 3U.S. Food & Drug Administration, Silver Spring, MD, 4OpenClinica, Waltham, MA

Introduction:
Standardized and efficient approaches to data capture are critical elements in clinical drug development. However, the abstraction of electronic health record data and entry into case report forms represents a major barrier for timely data capture, particularly where toxicities and safety must be monitored in near real time. This is especially relevant during a public health crisis such as COVID-19, where there is an urgent need to efficiently enroll a diverse population of subjects for rapid identification of safe and effective treatments. We leveraged the OneSource platform in the I-SPY COVID platform trial, in which integrations with local Electronic Health Record systems (EHRs) support automated capture of structured source data across multiple clinical centers.

Methods:
The OpenClinica (Waltham, MA) Electronic Data Capture (EDC) system was configured to support randomized subjects for 4 concurrent investigational agents and a control arm (remdesivir). A daily data entry electronic case report form (eCRF) of critical clinical data elements was defined that included the World Health Organization (WHO) Clinical Progression Scale and key Acute Respiratory Distress Syndrome (ARDS) clinical questions. A SMART on FHIR App was developed to launch within the EPIC Electronic Health Record (EHR) with support for source data capture of laboratory, concomitant medications, and demographic data using FHIR APIs. The solution allows research coordinators and investigators a one-click launch of the participant’s record in the EDC directly from the patient chart in the EHR, with single sign on and linkage between the records, eliminating need to separately log in to research applications. From the interface they can populate eCRFs with structured EHR data via FHIR by pushing a button, and allows review and/or manually enter data into eCRFs for which structured EHR data is not available.

Results:
The trial launched 30 July 2020, and as of 27 July 2021, 1,461 subjects had enrolled, with 742 randomized subjects receiving investigational agents. There were a further 719 subjects followed in the observational or RWD cohort. One of the I-SPY COVID sites with the highest I-SPY COVID accrual rates showed a 22% data entry error rate for D Dimer 3 and 4.5% error rate for lymphocyte count due to incorrect reference range entry for a subset of participants. This site is accelerating OneSource production implementation for August 2021. At UCSF, manual laboratory daily data entry for 5 subjects on average took 60 minutes, post OneSource implementation showed the same laboratory result data entered within 5 minutes through the EHR integration.

<table>
<thead>
<tr>
<th>Laboratory Test</th>
<th>Number of Errors</th>
<th>Total Records</th>
<th>Percent of Issues</th>
</tr>
</thead>
<tbody>
<tr>
<td>Absolute Neutrophil Count</td>
<td>45</td>
<td>10026</td>
<td>0.4%</td>
</tr>
<tr>
<td>D Dimer 3</td>
<td>1285</td>
<td>5836</td>
<td>22.0%</td>
</tr>
<tr>
<td>Lymphocyte Count</td>
<td>549</td>
<td>12325</td>
<td>4.5%</td>
</tr>
<tr>
<td>Platelets</td>
<td>95</td>
<td>18212</td>
<td>0.5%</td>
</tr>
<tr>
<td>WBC Counts</td>
<td>85</td>
<td>18289</td>
<td>0.5%</td>
</tr>
</tbody>
</table>

Table 1: Manual data entry errors for laboratory results for one of the I-SPY COVID sites over an 8-month period prior to OneSource implementation. OneSource capture prevents such errors. Figure 1: Efficiencies in laboratory result data capture by manual entry into the trial EDC vs OneSource EHR integration for 5 participants

Discussion:
Conducting clinical trials during a pandemic with limited research support staff requires streamlined approaches to data capture to rapidly assess drug efficacy and toxicities. This automated and EHR-integrated approach removes potential barriers and encourages adoption of the trial at hospitals with limited research infrastructure or where clinical research is not routinely performed.
Impact of Coronavirus-2019 on Pediatric and Adult Heart Transplantation Waitlist Activity and Mortality in the United States: A Descriptive Approach

Awais Ashfaq MD1, Geoffrey M. Gray PhD2, Jennifer Carapelluci3, Ernest K. Amankwah PhD4, Luis M. Ahumada PhD2, Mohamed Rehman MD2,5, Michael Puchalski MD6, Andrew Smith MD7, James A. Quintessenza MD1, Alfred Asante-Korang MD3

1Cardiovascular Surgery, Heart Institute, Johns Hopkins All Children’s Hospital, St. Petersburg, FL; 2Center for Pediatric Data Science and Analytic Methodology, Johns Hopkins All Children’s Hospital, St. Petersburg, FL; 3Heart Transplantation, Cardiomyopathy and Heart Failure, Heart Institute, Johns Hopkins All Children’s Hospital, St. Petersburg, FL; 4Epidemiology and Biostatistics, Johns Hopkins All Children’s Hospital, St. Petersburg, FL; 5Department of Anesthesia and Pain Medicine, Johns Hopkins All Children’s Hospital, St. Petersburg, FL; 6Division of Cardiology, Heart Institute, Johns Hopkins All Children’s Hospital, St. Petersburg, FL; 7Division of Cardiac Critical Care, Heart Institute, Johns Hopkins All Children’s Hospital, St. Petersburg, FL.

Abstract
This work provides a descriptive approach to characterizing the effects of the COVID-19 pandemic on the heart transplant community, with attention given to pediatric transplants. Transplants are found to increase in the adult community, while remaining consistent in the pediatric community. Mortality remains similar in pediatrics, while decreasing in adults.

Introduction
Transplant centers saw a substantial reduction in deceased donor solid organ transplantation since the beginning of the coronavirus 2019 (COVID-19) pandemic in the United States.1 There is limited data on the impact of COVID-19 on pediatric heart transplant volume and variation in transplant practices. We hypothesized that pediatric heart transplant activity decreased during COVID-19 with associated increased waitlist mortality.

Methods
The United Network for Organ Sharing (UNOS) database was used to identify patients at the time of listing for heart transplant from 2017–2020.2 Patients were categorized as pediatric (<18 years) or adult (>18 years) and as pre-COVID (2017-2019) or post-COVID (2020). Regional and statewide data were taken from United States Census Bureau. CovidActNow project was used to obtain COVID-19 mortality rates.

Results
The average number of pediatric transplants (n=39 per month) did not change significantly during 2020 (data not shown). However, there was a decline in the first quarter of 2020 followed by an increase. Absolute pediatric waitlist mortality decreased from 5.31 to 4.73. In adults, there was an increase in the number of transplants from 242 to 266. Adult waitlist mortality had a larger decrease from 18.44 to 15.70, with an increase in female mortality of 7%. Regional and demographics differences in mortality were observed and will be discussed further.

Conclusion
Pediatric heart transplant volume declined in early 2020 followed by a later increase, while adult transplant volume increased all year round. Although, overall pediatric waitlist mortality decreased, female waitlist mortality increased for both adults and pediatrics. Regional differences in waitlist mortality were observed for both pediatrics and adults. Future studies are needed to understand this initial correlation and to determine the impact of COVID-19 on heart transplant recipients.

References
Ensuring Covid patient end-of-life goals documented in an eMolst are honored through a novel and scalable health information exchange

Jonathan Austrian MD 1,2, Alex Low1, Victoria Javier RN1, Christine Wilkins PhD, LCSW2
1Department of Medicine, NYU Langone Health, New York, New York
2Medical Center Information Technology, NYU Langone Health, New York, New York

Introduction
Clinicians used Medical Orders for Life Sustaining Therapy (MOLST) as a means of for documenting healthcare preferences for individuals with serious illness and advanced frailty. The eMOLST is a digital version that is created and incorporated into the New York State eMolst registry. Health care systems in New York City worked to rapidly incorporate eMOLST to support advance care planning during the Covid crisis. Despite its importance, significant barriers to incorporating the eMOLST into patient care workflows have limited its potential. In order to determine if patients had eMOLSTs generated outside the health system, the clinician had to proactively search a separate eMOLST registry based on patient demographics. The objective of our work was to enhance the visibility of eMOLST status from within the electronic health record (EHR) regardless of where the eMolst was generated and streamline the review of those eMOLSTs.

Methods
In response the Covid-19 pandemic, our health system partnered with the state eMOLST registry, the health information exchange (HIE) (Healthix), and our EHR vendor (Epic Systems). We leveraged the HIE’s existing patient matching service. To ensure the availability of eMOLST status in real time, the HIE updated its clinical event notifications service to include the eMOLST number, if it exists, for all new or existing patients at our institution. Finally, we devised a new workflow within the EHR to promote awareness of a patient’s eMOLST status and ensure the content was readily available to clinicians. (Fig 1).

Fig. 1 Emolst Clinical Workflow

Conclusions
We successfully implemented this technical workflow on May 4, 2021. As a result of this partnership, our academic health system identified an additional 1,557 patients with eMOLSTs generated by other institutions, representing 11.5% more patients with eMOLSTs than previously known.

The Covid pandemic highlighted the importance of assessing goals of care and ensuring patients’ preferences were available at the point of care. Governmental registries can help overcome patient data siloes by centralizing the storage of this important information. We describe a clinical and technical workflow that enables care team members to visualize a patient’s eMOLST status directly in their clinical workflow using common vendors and processes that are readily scalable to other institutions.
A Machine Learning Approach to Identify EHR-based Cohorts of Pain Sensitivity for Genetic Analysis

Amelia J Averitt, MPH MA PhD1, Nilanjana Banerjee, PhD1, Michael N. Cantor, MD MA1
1Regeneron Genetics Center, Clinical Informatics, Tarrytown, NY, USA

INTRODUCTION. Among the many barriers to effective pain management is the variability in how patients respond to noxious stimuli, or response heterogeneity. The extent of these inter-individual differences is known as pain sensitivity. Patients that demonstrate an increased response to stimuli have high pain sensitivity (HPS); and patients that demonstrate a decreased response to stimuli have low pain sensitivity (LPS). Such classifications of the observable characteristics of pain sensitivity are phenotypes. Research demonstrates that an individual’s phenotype for pain sensitivity is attributable to a complex interaction of environmental factors and genetic variants. As such, genetic analyses that identify causal variants are a crucial tool for our understanding of pain sensitivity.

METHODS. This research presents a new machine learning approach to identify cohorts of high and low pain sensitivity from the electronic health record (EHR) to support a genetic analysis of causal variants. Our approach leverages each patients’ unique pain score responses within a single clinical encounter. For each patient, a density distribution is created from their pain score responses. We hypothesize that this distribution characterizes a patients’ pain sensitivity. A K-means algorithm is applied to partition the densities into clusters. Cluster centroids are defined by the mean, normalized density of cluster members. During the learning procedure, patients are re-assigned to nearest clusters, with this distance quantified with the Kullback-Leibler (KL) Divergence. Because the KL divergence is not symmetric, our clustering algorithm uses a distance \(D_{KL}(P,Q) = 0.5KL(P||Q) + 0.5KL(Q||P)\). We applied our model to real-world clinical data from the Geisinger Health System (GHS) EHR, subset to patients with whole exome sequencing (N=145k). To resist confounding by indication, we selected patients that underwent a total hip arthroplasty (N=1,904) or a total knee arthroplasty (N=3,696) with at least 1 pain score (NEMSIS) after the surgery end-time but within the same surgical encounter. We learned three clusters and kept those with the highest (HPS) and lowest median density (LPS). Cluster densities were evaluated by the Silhouette coefficient and a simulation and permutation t-test to determine significance of differences in densities skewness. To better understand the discriminative features between the clusters, we fit an L2 regularized logistic regression using age, gender, race, ethnicity, and the 500 most common diagnosis codes. To examine the genetic underpinnings of pain sensitivity between the HPS and LPS cohorts, we performed a genome-wide association study (GWAS).

RESULTS. The LPS cluster included 3,460 (64.6%) patients, the HPS cluster included 780 (14.6%) patients (Figure 1). The Silhouette coefficient of the clusters is 0.336. The simulation permutation t-test demonstrates at skew of -0.375 in the HPS cluster and 0.670 in the LPS cluster (p<0.0001). The regression indicates HPS patients have increased odds for painful conditions, mental health disorders, and exposure to tobacco. There were no genome-wide significant hits (Table 1). However, the variants with the most significant associations implicate genes involved with pain sensitivity in neurons (FCGR2), abdominal pain (FARS2), spinal nerve ligation-induced neuropathic pain (MNX1), and chronic pain after surgery (SOX4).

DISCUSSION. The results indicate that our machine learning approach is able to identify clinically meaningful cohorts of pain sensitivity from the EHR. The GWAS suggests that cohorts likely also represent different genetic architectures. Such insights into the molecular basis of pain may facilitate the identification of new potential targets for drug discovery. However, this method is not compared with a gold standard, as there is neither a preferred methodology for this task nor an objective measure. Greater evaluations are planned for future work.

A Monte Carlo Estimation of the Narrow-Sense Heritability of COVID-19 Infection and Severity from AncestryDNA Survey Data

Amelia J. Averitt, MPH MA PhD1, Deepika Sharma, MHI1, Michael N. Cantor, MD MA1
1Regeneron Genetics Center, Clinical Informatics, Tarrytown, NY, USA

INTRODUCTION. Respiratory infectious diseases, like COVID-19, demonstrate a host genetic component that contributes to inter-individual differences.1 This contribution of the host’s genetics, in tandem with environmental factors, determines disease susceptibility and infection. Narrow-sense heritability ($h^2$) summarizes the average proportion of total phenotypic variance that is due to additive genetic factors, rather than the environment. Knowledge of the $h^2$ of COVID-19 may be helpful to guide public health intervention.

METHODS. This research presents a Monte Carlo (MC) estimation of the $h^2$ of COVID-19 infection and severity from AncestryDNA survey data. This data is from a private collaboration with AncestryDNA and Regeneron Pharmaceuticals and was collected online from volunteer respondents. The survey was given to AncestryDNA customers (N=210k) and aimed to assess each respondent’s COVID-19 infection status, exposure, risk factors, and symptoms. An MC method was applied to this data to stochastically estimate the parent’s phenotypic variance using parameters that are grounded in real-world epidemiologic data (Algorithm 1). Stochastic parameters include the rates of two-parent households ($\eta$); household co-infection ($\rho$)$^2$; and hospitalization among COVID-19 infected individuals ($\phi$)$^2$. $h^2$ was then estimated from the slope of a linear regression that models the simulated, mid-phenotypic value of the parents by the phenotypic value of the respondent (the proband). This research presents an application of the proposed methods to (i) simulated data, and (ii) the AncestryDNA survey data to estimate $h^2$ of COVID-19 infection ($h^2_{\text{inf}}$) and severity ($h^2_{\text{sev}}$).

Simulation. Proband and parent phenotypes were simulated such that the ‘true’ $h^2$ was hard-coded into the data. If 1 or 2 parents were infected, the data was then masked to only indicate if 1 or more was affected, which aligns with the uncertainty in the AncestryDNA data. The MC algorithm was applied to this masked, simulated data and an $h^2$ estimate was made. This experimental setup was repeated using varying ‘true’ $h^2$ values of 0.1, 0.3 and 0.5.

Application to AncestryDNA data. In the event that a survey respondent noted that “Parent(s)” were infected or hospitalized (a proxy for severity), this data was fed into the MC simulation to stochastically model the parent phenotype -- if 1 or 2 parents were infected ($p^1$) and hospitalized ($p^h$). Models for both phenotypes of interest, were adjusted for the age, gender, and ever-smoking status of the proband. Given the high variability in reported household co-infection, $\rho$, the MC simulation was repeated using three reported statistics to inform this parameter.$^3$\-5

RESULTS. When applied to simulated data, in which ground-truth $h^2$ is known, the method is able to recover $h^2$ with high accuracy (Table 1). The application to AncestryDNA survey data indicates that $h^2$ of COVID-19 infection susceptibility and ranges from 0.155 to 0.183 and severity ranges from 0.070 to 0.075 (Figure 1). These results suggest a moderate genetic contribution to COVID-19 infection susceptibility and a low genetic contribution for COVID-19 severity.

CONCLUSIONS. This research provides evidence that variability in COVID-19 infection is, in part, explained by inherited genetic factors. However, this approach does not account for recessive effects or gene-gene interaction and is reliant upon the estimated parameters. These results are limited by the heterogeneity of AncestryDNA. Further experimentation is necessary to understand the limitations and validity of this method.


---

Table 1. Mean MC estimates and 95% confidence intervals of $h^2$ when applied to masked, simulated data with target $h^2$’s of 0.1, 0.3, and 0.5.

<table>
<thead>
<tr>
<th>$h^2$ estimate</th>
<th>0.1</th>
<th>0.3</th>
<th>0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.093 [0.079, 0.129]</td>
<td>0.323 [0.302, 0.342]</td>
<td>0.479 [0.460, 0.498]</td>
<td></td>
</tr>
</tbody>
</table>

---

Figure 1. MC estimates of $h^2$ from AncestryDNA data, with varying parameters of household co-infection = 0.163, 0.170, and 0.530.
Dashboard for Machine Learning Models in Health Care
Wejdan H. Bagais, BS, Janusz Wojtusiak, PhD
Department of Health Administration and Policy, George Mason University, Fairfax, VA

Introduction
Presentation of machine learning (ML) models and their results plays an important role in analysts’ and clinicians’ understanding of predictions and consequently trust in the models. Visualization methods are among the best ways to explain the model performance as noted by Tonekaboni et al. who emphasize that carefully designed visualizations increase the clinicians’ understanding. While a considerable amount of literature has been published on explaining the performance of ML models, most studies focus on one measure or a specific ML method. This work summarizes the most important factors for evaluating any classification supervised ML model in one place using a dashboard.

Methods
Figure 1 shows a high-level design of the dashboard that is split into three main sections: statistical measures, features importance, and sensitivity analysis.

Statistical measures: This section visualizes model training and testing performance: AUC, accuracy, recall, precision, and f-score using a heatmap followed by the ROC curve with the best threshold. Then the confusion matrix is shown using stacked mosaic plots suggested by Raymaekers et al. paper to show the proportion of observations in each class by the squares’ size to indicate if the data are skewed. Finally, distribution of predicted values is displayed to show the model’s confidence levels.

Feature importance: This section visualizes the correlation, feature importance based on LASSO, random forest (embedded methods), and permutation (wrapper method), two learning curves (one based on the number of attributes and the other based on the number of observations). To avoid cluttering, the number of displayed attributes is limited to the 20 that score the highest.

Sensitivity analysis: This section visualizes the effects of a single attribute on the model using the technique described in Wojtusiak et al. study. After selecting an attribute, the dashboard shows its distribution, and mean predicted values for training, testing, and random data, as well as means of fixed values for the attribute to check its effect in the model. Finally, two scatter plots are displayed to compare the original predictions and predictions when the attribute values change slightly. For categorical attributes, the distribution of predicted values is displayed for each category instead of the mean.

Results
Medicare claims, COVID-19, and heat disease prediction data from UCI repository were used to test the visualization. For claims data, the purpose was to predict if the patient will be a high utilizer in the next six months, while COVID-19 data were used to predict covid-19. Figure 1 shows the dashboard for predicting high utilizers using claims patients.

Discussion
Visualizing the model results help both the analysts and the clinicians to evaluate the model results. Based on the problem, different ways of evaluation are needed. However, this dashboard visuals are applicable to any type of classification problem. The next step of this project is to visualize the difference between multiple models.

References
Use of Geomapping and Data-driven Personas to Target Intervention for At-risk Patients with Diabetes

Adarsha S Bajracharya, MD, MMSc1,2, Soussan Djamasbi PhD,MS,3 Gaayathri Sankar BA,3 Qiming Shi, MS1 Daniel J Amante PhD, MPH,1

1. Dept of Population and Quantitative Health Sciences, University of Massachusetts Medical School, Worcester, MA 2. Dept of Medicine, University of Massachusetts Medical School, Worcester, MA, Worcester, MA 3. User Experience and Decision Making Lab, Worcester Polytechnic Institute, Worcester, MA

Introduction:

Diabetes is a major health concern in the US with over 25 million Americans having type 2 diabetes (T2D). Adherence to guideline recommended care has shown to improve health outcomes in diabetic patients, including decreased rates of complications, less hospitalizations and improved quality of life. Yet, adherence to guideline recommended diabetes care plan is a challenge for many patients due to physical and socio-economic barriers. Persona development refers to the process of creating representations of patients via fictional yet realistic characters (personas). This has been shown to be important in identifying common needs, allowing for development of personalized interventions. We sought to use select demographic and clinical data of diabetic patients from an institutional electronic health record (EHR) system to geemap and create patient personas to identify groups of patients with common characteristics. This works allows us to better understand a diverse patient population and their corresponding needs.

Methods:

UMass Memorial Healthcare (UMMHC) is a large academic health care system caring for patients in central Massachusetts. Institutional patient data from its diabetes registry were used for geomapping the patient population and for cluster analysis to create data driven patient personas. Patient home addresses, including street address, city, state, and zip code, were extracted from EHR system and ArcGIS application was utilized to geocode and transform patient home addresses into latitude and longitude enabling spatial cluster analysis. Two-step clustering technique using SPSS was employed to analyze categorical EHR data to identify representative patient groups.

Results:

Among 53,675 deidentified patient records, analysis of the geocoded data revealed 10 zip codes with highest burden of disease. These 10 zip codes accounted for 48% of the patient population.

The two-step cluster analysis was run for this target population of 26,349 patients and three clusters of patients with important similarities and differences were identified. Common features included being white, obese, with moderate LACE+ comorbidity score, having access to phone and an email address (suggestive of access to Internet) but were not active patient portal users. Important differences noted in the 3 clusters were one cluster of younger, single females of Hispanic/Latinx ethnicity with state-sponsored insurance and two other clusters consisting of older married males not of Hispanic/Latinx ethnicity who had either Medicare or commercial insurance.

Conclusion:

Early results show that geomapping and cluster analysis can be used to identify different subgroups from a larger patient population and help to understand their group characteristics. In our work, we identified communities with highest burden of diabetes. Next steps include gaining further understanding of the barriers and facilitators to guideline-concordant care for each subgroup of patients within the population. This will help us better understand specific clinical and social determinants of health, allowing for tailored development of patient engagement strategies and support services that better meet the needs of the patients.
Electronic Health Record Use for Improving Patient Outcomes for School-Aged Children with Chronic Conditions

Christina Baker, MS, RN-BC¹, Figaro B. Loresto Jr., PhD, RN²
¹University of Colorado | Anschutz Medical Campus, Aurora, CO, USA, ²Children’s Hospital Colorado, Aurora, CO, USA

Introduction

Electronic health records (EHRs) have become the standard of communication and documentation for healthcare systems. Still, full functionality is not yet realized for health information exchange (HIE) between healthcare systems and community partners such as school nurses. One barrier to HIE includes misunderstanding of Federal guidelines, the Health Insurance Portability and Accountability Act of 1996 (HIPAA) and the Family Educational Rights and Privacy Act of 1974 (FERPA) (1). One pilot study gave EHR access to school nurses: results showed decreased hospitalizations pre and post school nurse EHR access for 33 school-aged pediatric asthma patients (2).

An innovative program at a Rocky Mountain area Children’s Hospital provides school nurses based in local urban area schools access to a view-only version of the hospital-based EHR, EpicCare Link (Epic Systems, Verona, WI). With Internet access, the school nurses log into EpicCare Link and view the student's chart, including the longitudinal plan of care, current medications, school health care plans, immunizations, and encounter notes. Access process includes school district and hospital signing a Health Information Sharing Agreement. School nurses complete an online EHR training, sign a Security User Agreement, and complete an in-person privacy and technical training.

Objective/Methods

The study aimed to analyze the effect of school nurse access to medical records in a hospital-based EHR on patient outcomes. We hypothesized that EHR access would decrease emergency department (E.D.) visits and hospital inpatient admissions and that EHR access would be utilized more for patients with chronic conditions.

This retrospective secondary data analysis was conducted in 2020 using EHR data six months pre- and post-school nurse access to students’ hospital-based EHR. The main outcome measures were the E.D. visits and hospital admissions. Descriptive analysis was conducted on demographic variables. ED visits and hospital admissions were not normally distributed by the Shapiro-Wilks test (W = 0.52, p < 0.0001; W = 0.52, p < 0.0001). Unadjusted matched Wilcoxon Rank Sum tests were conducted to compare outcomes between the pre-and post-access time periods for each subject. A multivariate mixed Poisson regression was conducted to estimate the impact of EHR access for school nurses on outcomes adjusting for covariates listed in the demographic and covariate section.

Results

For the sample of 336 students in the study, there was a 33% decrease in E.D. visits from 190 visits before access to 126 ED visits after access. Hospital admissions decreased by 44% from 176 hospital admissions before access to 99 hospital admissions after access. Poisson Regression analysis found a significant (p<.001) 34% reduction in E.D. visits and 44% reduction (p<.01) in inpatient visits post-access compared to pre-access. EHR access was utilized by school nurses more for patients with higher complexity and chronic conditions.

Discussion/Conclusion

HIE between hospital systems and community partners was established. School nurse access to medical records through a hospital-based EHR may improve patient outcomes with access to real-time, accurate medical records. Further research of feasibility, improved functionality, and confounding variables is warranted.

References

Planning Change Using the ADKAR Model: Establishing a Data Governance Program in the VHA Office of Community Care

Gina Baker, CCP, Francine Sandrow, MD, Katharine Miller, MHA/MPM, Jianji Yang, PhD, Darby McKnight, MBA, Natalie Ngu, MS
Department of Veterans Affairs, Veterans Health Administration, Office of Community Care

Introduction
Veterans Health Administration (VHA) provides access to timely, high-quality, cost-efficient, and well-coordinated community care for Veterans and VA beneficiaries based on certain conditions and eligibility requirements. The VHA Office of Community Care (OCC) is the national office in charge of community care delivery and revenue operations. Data management and analytics are critical to support OCC administration, planning, and oversight functions.

The evolution of VA Community Care operations and implementation of third-party software to manage referrals, claims and payments, has contributed to OCC data systems becoming overly complex and disjointed. Subsequently, OCC established a Data Governance (DG) Program to manage data as a strategic asset and formalize behavior and control over how data assets are managed. The DG program will establish Data Stewardship practices from data capture to disposition, to ensure data is consistent and of the highest quality with the goal of improving care quality and Veteran satisfaction informed by data analytics.

Methods
The DG program imposes significant changes on staff and current processes. OCC has executed a methodical way to manage change, using the Prosci ADKAR Model to guide staff through the transition. ADKAR stands for the 5 elements for a successful change: Awareness of the need for change; Desire to support and participate in the change; Knowledge of how to change; Ability to implement required skills and behaviors; and Reinforcement to sustain the change.

A DG project team was assembled by the OCC Chief Health Informatics Officer and the change management component is led by a Certified Change Practitioner. The team conducted a baseline assessment using Prosci tools. Through interviews and observations, project health and the Impact Index for 33 groups of staff were assessed in line with the 5 elements. Based on the insights gained from the assessment, the team created an organizational change management plan with mitigation strategies tailored for each group to address the risk and barriers. In accordance with the organizational change management plan, the following plans will be developed in support of ADKAR: sponsor roadmap, coaching plan, communications plan, resistance management plan and training plan.

Results
After analysis of the project and group impact results, a change management plan based on the ADKAR Model is created with strategies for each step in the lifecycle of the project (Figure 1).

Conclusion
Change management plays a crucial role in project success, especially for high impact projects like the DG Program. Using the ADKAR change framework, leadership has a better understanding of the risks and resistance to the adoption of the Data Governance strategy in OCC and is developing plans to mitigate the risks and ensured project success.

Figure 1. Change management plan for each step of the project lifecycle outlining the strategy for each element of ADKAR Model.
Evaluating HL7 FHIR Resources for Sharing Research Consent Data

M Katie Banasiewicz; Mark McEver; Douglas Conway, MS; Alex C Cheng, PhD, MEM; Colleen Lawrence, PhD; Leah Dunkel, MPH; Paul A Harris, PhD; Stephany N Duda, PhD
Vanderbilt University Medical Center, Nashville, TN, USA

Introduction
The HL7 Fast Healthcare Interoperability Resources® (FHIR®) specification was originally developed to facilitate the exchange of clinical care data from the electronic health record, but is increasingly used to share research data. We evaluated the capacity of FHIR Release 4 to represent the elements of informed consent for research.

Methods
With researcher input, we identified the data and metadata elements needed to describe different types of research informed consent, reviewed HL7 FHIR Resources for appropriate mappings, and identified gaps.

Results
We identified 21 data elements to represent research consent and associated documents (e.g., child assent), of which 17 (81%) mapped to existing FHIR Resources (Table). No FHIR elements aligned with data for consent version, language or mode of administration (e.g., in-person, mail, electronic), or copy of the signature image (for eConsent).

Table. Research consent data elements and their mappings to HL7 FHIR Resources.

<table>
<thead>
<tr>
<th>Data Element(s)</th>
<th>FHIR Resource</th>
<th>FHIR Element(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Required FHIR metadata</td>
<td>Consent</td>
<td>status (active/inactive/entered-in-error), scope (research), category (patient/release of information)</td>
</tr>
<tr>
<td>Researcher consenting patient</td>
<td>Consent</td>
<td>performer</td>
</tr>
<tr>
<td>Date &amp; time (could be when researcher signs)</td>
<td>Consent</td>
<td>dateTime</td>
</tr>
<tr>
<td>Person to whom consent applies</td>
<td>Consent</td>
<td>patient</td>
</tr>
<tr>
<td>Research study site</td>
<td>Consent</td>
<td>organization</td>
</tr>
<tr>
<td>Person signing the consent</td>
<td>Consent</td>
<td>verifiedWith</td>
</tr>
<tr>
<td>Date &amp; time consent signed by patient/representative</td>
<td>Consent</td>
<td>verifiedDate</td>
</tr>
<tr>
<td>Summary or full text of the informed consent</td>
<td>Consent</td>
<td>text</td>
</tr>
<tr>
<td>Link to the associated DocumentReference</td>
<td>Consent</td>
<td>sourceReference</td>
</tr>
<tr>
<td>Person who signed document (for assoc. documents)</td>
<td>DocumentRef.</td>
<td>authenticator</td>
</tr>
<tr>
<td>One or more files in the form of Attachment Resources (e.g., child assent, consent short form, attestations, optional substudies if not in main consent)</td>
<td>DocumentRef.</td>
<td>content</td>
</tr>
<tr>
<td>Attachment content (e.g., PDF)</td>
<td>Attachment</td>
<td>attachment</td>
</tr>
<tr>
<td>Document descriptors (language, title/description)</td>
<td>Attachment</td>
<td>language, title</td>
</tr>
<tr>
<td>Date &amp; time (could be signing time)</td>
<td>Attachment</td>
<td>creation</td>
</tr>
</tbody>
</table>

Discussion
FHIR Resources, particularly Consent and DocumentReference, currently cannot capture the full data structure of research consents, but hold promise for representing research consents via extensions or future versions. Limitations include our focus on US-based research consents or unrecognized consent fields. We are developing a proof-of-concept Consent FHIR payload module built into the REDCap data collection platform to implement this mapping.

Funding: This work is supported by NIH/NLM contract #75N97019P00279 and NIH/NCATS award U24TR001579.

References
Patient Perspectives on ‘FH Family Share,’ a Tool to Increase Uptake of Cascade Testing for Familial Hypercholesterolemia

Hana Bangash¹, MBBS, Ahmed Makkawy², BSc, Justin H. Gundelach¹, MSc, Alexandra Miller¹, BSc, Aaron Leppin³, MD, Iftikhar J. Kullo¹*, MD.

¹Department of Cardiovascular Medicine, Mayo Clinic, Rochester, MN; ²User Experience Research, Saharafox Creative Agency, Rochester, MN; ³Health Care Delivery Research, Mayo Clinic, Rochester, MN, USA.

Introduction

Familial hypercholesterolemia (FH), a prevalent genomic disorder, significantly increases risk of early onset coronary heart disease. Cascade testing for FH is a cost effective method of identifying at-risk family members of a diagnosed proband, however it has low uptake in the U.S.¹ Digital tools may increase uptake by facilitating the sharing of FH-related risk information between probands and family members.

Methods

‘FH Family Share’ is a web-based tool designed to promote cascade testing for FH by enabling probands to share their diagnosis with family members via templated ‘letter to family members’ emails. To evaluate FH Family Share we conducted usability testing with FH patients over a 3-month period. Patients with a confirmed pathogenic/likely pathogenic variant were asked to participate. Due to Covid-19, sessions were conducted virtually, lasted 1 hour, were audio recorded and transcribed. Patients navigated a prototype of the tool, applied the ‘Think Aloud’ usability technique and shared insights on interface design, content and user experience. Patients also completed a survey.

Results

Nine patients participated in the usability testing sessions. Feedback highlighted the need to address the implications of FH for children in the tool, including age of screening and treatment, and implications for FH patients who were family planning. Children and grandchildren were identified as the greatest motivators for patients to get tested and share their diagnosis. Feedback on interface design included need for larger, clearer font and making the content easier to read. Patient feedback led to iterative refinement of the tool. Survey results are summarized in Table 1.

Table 1: Satisfaction survey responses from FH patients (n=9).

<table>
<thead>
<tr>
<th>Survey Questions</th>
<th>Patient Responses (n(%))</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Overall, the information that you were asked to assess within the FH Family Share website was:</td>
<td></td>
</tr>
<tr>
<td>o Very easy to find</td>
<td>5 (55.6)</td>
</tr>
<tr>
<td>o Somewhat easy to find</td>
<td>4 (44.4)</td>
</tr>
<tr>
<td>2. Overall, the information that you found within the FH Family Share website was:</td>
<td></td>
</tr>
<tr>
<td>o Very easy to understand</td>
<td>7 (77.8)</td>
</tr>
<tr>
<td>o Somewhat easy to understand</td>
<td>2 (22.2)</td>
</tr>
<tr>
<td>3. How will the FH Family Share website impact patient care/follow-up care?</td>
<td></td>
</tr>
<tr>
<td>o Significantly improve</td>
<td>3 (33.3)</td>
</tr>
<tr>
<td>o Somewhat improve</td>
<td>5 (55.6)</td>
</tr>
<tr>
<td>o Neither improve nor worsen</td>
<td>1 (11.1)</td>
</tr>
<tr>
<td>4. The FH Family Share website will make it easier for patients like myself to understand/Share an FH diagnosis:</td>
<td></td>
</tr>
<tr>
<td>o Completely Agree</td>
<td>5 (55.6)</td>
</tr>
<tr>
<td>o Agree</td>
<td>4 (44.4)</td>
</tr>
<tr>
<td>5. As a patient, I would be most likely to choose only one option:</td>
<td></td>
</tr>
<tr>
<td>o Use the Learn modules as a knowledge resource</td>
<td>2 (22.2)</td>
</tr>
<tr>
<td>o Build a family tree using AboutMe</td>
<td>0 (0.0)</td>
</tr>
<tr>
<td>o Calculate risk of a heart attack</td>
<td>3 (33.3)</td>
</tr>
<tr>
<td>o Send a letter to family members</td>
<td>2 (22.2)</td>
</tr>
<tr>
<td>o Use website information to discuss FH with family members</td>
<td>2 (22.2)</td>
</tr>
<tr>
<td>6. Do you find the FH Family Share website Figures/Images/Diagrams useful?</td>
<td></td>
</tr>
<tr>
<td>o Yes</td>
<td>8 (88.9)</td>
</tr>
<tr>
<td>o No</td>
<td>1 (11.1)</td>
</tr>
<tr>
<td>7. Is the FH Family Share website a resource worth returning to?</td>
<td></td>
</tr>
<tr>
<td>o Yes</td>
<td>9 (100)</td>
</tr>
<tr>
<td>o No</td>
<td>0 (0.0)</td>
</tr>
</tbody>
</table>

Conclusion

FH Family Share, an innovative digital tool, has the potential to promote cascade testing and reduce FH-related morbidity and mortality. FH Family Share will be further evaluated in a pilot implementation study.

References

Assessing the Use of Prescription Drugs in Obese Respondents in the National Health and Nutrition Examination Survey

Laura A. Barrett, MS¹, Aiwen Xing¹, Elizabeth Steidley¹, Terrance J. Adam, RPh, MD, PhD², Rui Zhang, PhD², Zhe He, PhD¹*  
¹Florida State University, Tallahassee, FL; ²University of Minnesota, Minneapolis, MN

Introduction

Obesity is a complex, multifactorial, and largely preventable disease commonly defined as having an excess of body weight for an individual’s height. As a major health and economic crisis affecting the modern world, much progress has been made in identifying and developing strategies for preventing and treating obesity. It is important to understand the patterns of RXD use among people with obesity, who often have other chronic conditions, in order to inform both clinical practice and research. In this project, we aim to gain an in-depth understanding of both the relationship between obesity and RXD use, as well as the correlations between specific RXD and DS use in the obese population using National Health and Nutrition Examination Survey. We would like to understand: 1) what are the correlations between reported dietary supplement use, reported prescription drug use, and demographics? 2) Are prescription drugs or dietary supplements used more by the obese population than in the non-obese population? 3) Can demographic variables and obesity status predict prescription drug or dietary supplements use?

Methods

Demographic, physical examination, prescription drug (RXD) and dietary supplement (DS) use, and health insurance information were extracted from NHANES for survey years 2003 – 2014 (6 survey cycles). We used the sample weight provided by NHANES to combine the samples from different survey cycles. We conducted multivariate logistic regression analyses, performed two separate logistic regression analyses and evaluated different machine learning models to answer our research questions.

Results

The obese group has a higher reported use of RXD. Use of CV agents and metabolic agents was higher in the obese group. The highest average number of RXDs used can be found among people with BMI from 63-67 kg/m². Figure 1 illustrates the correlation between the number of RXDs/DSs used and BMI. People take higher numbers of RXDs than DSs. The largest maximum number of DSs used (24) can be found among people with BMI from 18 to 22 kg/m², the healthiest group, compared with other BMI groups. Regarding the prediction results, we discovered that adding the extra predictor (of RXD/DS use) did not make a significant contribution to the predictions for RXD or DS use. For RXD prediction, insurance status, DS use, and age were the top three important features.

Conclusions

We found that there are demographics that show higher DS and RXD use; and that obesity can be a factor in the types of RXD that are used. Further knowledge on the association between obesity, DS, and RXD can inform patient education with the help of informatics tools such as data dashboards.

References

Asthmatic Patient Portal Messaging and Pollution

Marily Barron, BS, Matthew D. Zaragoza-Watkins, PhD, Melissa L. McPheeters, PhD MPH, S. Trent Rosenbloom, MD MPH FACMI FAMIA
Vanderbilt University, Nashville, TN, USA

Introduction

Air pollution is the most common allergen to trigger asthma exacerbations or asthma attacks1 and this is well established in scientific literature2. Self-management is a hallmark of asthma treatment, and includes symptom tracking, trigger awareness, medications, and health-seeking when risk of exacerbation increases.

This study will evaluate disparities in health seeking behavior among adult asthmatic patients during spikes in ambient air pollution. Broadly defined, health seeking behaviors are attempts to access health-related knowledge and resources, including healthcare providers and health systems, with the goal of improving health status3. This study will measure health seeking behavior using the frequency of patient portal messages sent from patients to physicians. The use of patient portals to communicate with healthcare providers is a potentially important form of health seeking behavior4, which may vary in diverse populations. We are not aware of prior research evaluating the health seeking behavior of asthma patients through patient portals when exposed to unfavorable environmental circumstances.

Methods

This retrospective study will be performed at Vanderbilt University Medical Center (VUMC), a large academic medical center located in Nashville Tennessee that serves primary-care and referral patients from across the Southeast. Subjects will include all adult asthmatic outpatients who visited a VUMC clinical from 2012 to 2015. Portal messaging of included patients will be measured from their initial visit until 2017, providing at least 2 years of longitudinal data for each outpatient. We follow, a recent study of adult asthmatic patients in Chicago4, identifying asthma patients as individuals who are assigned an ICD9 code for asthma (493.*) in their first observed visit to a VUMC clinic and are assigned either the asthma ICD9 code or a medication related to asthma on a subsequent visit5.

Data for this study will be collected from VUMC’s Electronic Health Record (EHR), My Health at Vanderbilt (MHAV), and Vanderbilt’s pollution data bank. Patient message counts will be collected from MHAV, VUMC’s patient portal. Medical (presence of asthma) and demographic (age, race and residence) variables will be extracted from the EHR. Ambient concentrations of various air pollutants (e.g., fine particulate matter) will be collected from the Vanderbilt Air Pollution Data Bank. This ensemble will be used to generate a categorical variable, Air Quality Index (AQI), aggregated by Zip-code-date.

To model the relationship between health seeking behavior and air quality, we will estimate a linear probability model via a panel fixed effects research design. Fixed-effects regression avoids potential confounding due to unobserved socioeconomic factors that covary with AQI and affect health seeking behavior at the level of Zip code or date. These statistical models will be used to determine if there is a significant increase in messaging frequency during episodes of high air pollution levels compared to those with low air pollution levels. Using EHR data, we will also estimate models that test whether the magnitude of this effect is mediated by socioeconomic factors such as race and income.

Conclusion

Our findings may reflect differences in the hesitancy of different demographic groups to use the patient portal to communicate with their clinician, differences in familiarity with technology or differences in asthma exacerbation.

References

Towards effect estimation of COVID-19 Non-pharmaceutical Interventions

Vesna Barros, MS¹, Victor Akinwande, MS², Itay Manes, BS¹, Osnat Bar-Shira, PhD¹,
Celia Cintas, PhD², Yishai Shimoni, PhD¹, Michal Rosen-Zvi, PhD¹
¹ IBM Research, Mount Carmel Haifa, Israel; ²IBM Research Africa, Nairobi, Kenya;

Introduction There still remains uncertainty about the relative effectiveness of multiple Non-pharmaceutical Interventions (NPIs) applied by different countries against COVID-19 transmission. We aim to estimate the effect of these NPIs on reducing the spread of the pandemic using a causal inference approach.

Methods We used Google mobility data¹ and a subset of socioeconomic and health covariates from World Development Indicators² as covariates in our causal analysis. We obtained comprehensive NPI data provided in WNTRAC¹ and the number of COVID-19 confirmed cases were extracted from the World Health Organization³.

To estimate the causal effect of a given NPI, we defined the outcome of the causal model \( Y \) as the reproduction number’s \( R_t \) slope two weeks after the NPI was imposed, where \( R_t \), i.e., the rate by which the pandemic spreads, was calculated using EpiEstim². Treatment was defined based on the imposition \( A = 1 \) or lifting \( A = 0 \) of the NPI. Using this problem setup, Inverse Probability Weighting (IPW)³ and Adversarial Balancing (AB)⁴ were used to calculate the weighted population outcome for each subgroup, stratified by treatment assignment. Using the weighted population, the average treatment effect (ATE), i.e., \( E[Y_{a=1}] - E[Y_{a=0}] \) was calculated, where it was assumed that the null hypothesis is that not imposing any NPIs would result in an ATE=0.

Preliminary Results We selected the ten most frequently imposed NPIs across 115 countries and 50 US states from the beginning of the pandemic until May 2021. For each NPI, we adjusted for confounding bias by balancing the distribution of the covariates between the groups with that NPI imposed and not imposed. Based on the evaluation framework proposed by Shimoni et al.⁵ we observed improved balancing after applying IPW for NPIs corresponding to Confinement (C), School Closure (S), Mass gatherings (MG), Mask wearing (MW), and Work restrictions (W). The population average treatment effect of these NPIs on \( R_t \) slope are reported in Table 1 with the 95% CI. We observed an overall null effect of these NPIs when multiple countries and US states are aggregated.

Discussion Given the heterogeneity of different countries, finding valid causal effects is challenging. Not only the covariates are noisy and uncertain, but unobserved factors may also confound the results. Additionally, the outcome, \( R_t \) slope, is itself an estimate with some degree of uncertainty. We recommend that any appropriate data-driven NPI effectiveness model must be accompanied by extensive validation and sensitivity analyses. As future work, we will do an individual evaluation for the US states only, and redefine the treatment and control groups of our causal model.

References


¹https://www.google.com/covid19/mobility
³https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports
Text Mining on COVID-19 Publications  
Mrigendra M. Bastola¹, MD, MS, Paul Fontelo¹, MD, MPH.  
¹NIH/NLM/LHC, Bethesda, MD, USA.

Abstract:  
The unprecedented number of publications on COVID-19 has created a challenge for physicians, researchers and consumers to get reliable health information. We deployed text mining tools (PyTorch) on selected PubMed articles to answer context-specific COVID-19 related questions, aiming to develop an askMEDLINE like tool.

Introduction:  
COVID-19 has generated unprecedented numbers of publications, challenging health professionals and consumers seeking trusted, reliable information from the internet. Text mining tools that accurately mine a reliable publications portal (e.g., PubMed) can better answer COVID-19 questions. This study aims to develop a text mining tool similar to askMEDLINE using Natural Language Processing to find PubMed resources.¹

Methods:  
Selected publications were searched in PubMed and resulting PubMed IDs were recorded. Using PyTorch, queries were modeled after the PICO (patient, intervention, comparison, outcome) for context-specific clinical questions about COVID-19 (outcome, prognosis, laboratory tests, age, and comorbidities). Query methods included CORD-19 and BERT models. Word vector analyses were performed, heat maps using RoBERTa models were generated. Search sequences were generated in PyTorch, using pre-trained models from a “transformers” model.

Results:  
Text mining tools deployed retrieved relevant material from selected PubMed publications. The results of the query varied for PICO query type and vocabulary (Figure 1). Results showed high association of sample query outputs to commonly used COVID-19 medical terms and phrases (0.7 to 1.0) (Figure 1 & Table 1). Overall, vector similarity analysis on the query output showed 0.67 similarities within words with similar meanings and 0.49 for words with different meanings. The average cosine distance was 0.03 between search words and output.

Conclusion:  
Text mining methods may provide a useful tool for healthcare givers and consumers looking for reliable information on COVID-19. Natural language queries from healthcare consumers, supported by healthcare givers, could improve access to their health information needs and make better medical decisions. This method can also be applied to other reliable sources (e.g., medRxiv), while also addressing the quality of published evidence.

Acknowledgments: This research was supported by the Intramural Research Program of the NLM/NIH.

References:  
A Mobile app to Stratify Deterioration Risk at Admission for COVID-19 Patients Using the Rothman Index

Joseph Beals IV, Ph.D.¹, Jaime J. Barnes D.O.², Daniel J. Durand, M.D.³, Joan M. Rimar, D.N.Sc.⁴, Thomas J. Donohue, M.D.⁴, S. Maftuz Hoq, M.D.⁵, Kathy W. Belk, B.A.¹, Alpesh N. Amin, M.D.⁶, Michael J. Rothman, Ph.D.¹

Affiliations: (1) PeraHealth, Inc., Charlotte, NC, (2) Sinai Hospital of Baltimore, Department of Medicine, Baltimore, MD, (3) LifeBridge Health, Baltimore, MD, (4) Yale New Haven Hospital, New Haven, CT, (5) Bridgeport Hospital, Bridgeport, CT, (6) The University of California, Irvine Medical Center, Orange, CA

INTRODUCTION
Hospitalized SARS CoV-2 coronavirus disease (COVID-19) patients are susceptible to a dysregulated immune system response giving rise to significant deterioration, days after admission. Among COVID-19 patients, reports indicate that age and comorbidity correlate with mortality risk. Our work evaluates the relationship between impairment and subsequent in-hospital deterioration by explicitly measuring physiologic acuity at admission using the Rothman Index (RI), an extensively validated, commercially available patient condition score (PeraHealth, Inc.; Charlotte NC). We evaluate the ability of the RI model at the time of hospital admission to stratify patients into high or low risk of subsequent severe deterioration with an emphasis on opportunities to support risk stratification in the COVID-19 population.

METHODS
Data - We analyzed data from LifeBridge Health System’s Sinai Hospital of Baltimore, in Baltimore, MD and from three Yale New Haven Health System (YNHHS) hospitals in New Haven, Bridgeport, and Greenwich CT. These include urban and suburban community hospitals and an academic medical center. The data included 3,499 COVID-19 patients and 14,658 non-COVID-19 patients from January 2020 – June 2020.

RI at Admission as a Predictor of Risk
To evaluate differences in deterioration risk between COVID-19 and non-COVID-19 patients, we compared RI at the time of inpatient admission with in-hospital mortality rates for COVID-19 and non-COVID-19 populations. We validated the use of the RI with COVID-19 patients by using a single-variable logistic regression model to compare the first RI score in predicting in-hospital mortality.

RESULTS
RI at Admission as a Predictor of Risk
Evaluation of in-hospital mortality as a function of acuity as measured by the RI at admission reveals that COVID-19 patients have a significantly higher mortality risk than non-COVID-19 patients for a given acuity. For both COVID-19 and non-COVID-19 patients, performance at all hospitals using the initial RI on admission ranges from good to excellent for discriminating in-hospital mortality (AUC of 0.81–0.84 for COVID-19 and 0.90–0.92 for non-COVID-19). The RI model is found to be well calibrated.

Operating Thresholds for Risk Stratification
Fewer than a third of COVID-19 patients meet the high-risk RI criteria (RI<50); they have a 34–45% mortality rate. Approximately half the COVID-19 population meet the low-risk RI criteria (RI>70); they have a mortality rate of 2–5%.

DISCUSSION
A direct measure of acuity is an effective predictor of physiological susceptibility to decline as shown in reported AUCs. Importantly, we also see that for a given level of acuity at admission, as measured by the RI, mortality rates for COVID-19 patients are higher than for the non-COVID-19 medical population; this difference holds even when comparing COVID-19 to a sub-population of non-COVID-19 medical patients with respiratory-related diagnoses. These findings imply that a clinician’s sense of how to treat a COVID-19 patient may be led astray by how he or she would treat a non-COVID-19 patient. Especially during surge situations, an objective measure may help ensure that resources, supplies, and bed capacity are allocated efficiently. These findings are also important for high-risk patients who may benefit from closer monitoring or more intensive therapies, and for low-risk patients, some of whom may not need high levels of care, or indeed hospitalization at all. In conjunction with the Biomedical Advanced Research and Development Authority we have developed a mobile app which we are currently testing for risk stratification in the Emergency Department to assist clinicians in admission and level-of-care decisions.
A Process for Multi-Stakeholder Governance of Patient Data Releases

Douglas S. Bell, MD PhD1,2,3, Marianne Zachariah1, Amanda L. Do1, Carina V. Hampp1, Karen V. Lopez1, Michael Pfeffer, MD1,2,3

1UCLA Clinical & Translational Science Institute, Los Angeles, CA; 2Department of Medicine, David Geffen School of Medicine at UCLA, Los Angeles, CA; 3UCLA Health, Division of Information Services and Solutions, Los Angeles, CA.

Introduction

Data derived from patient care is needed for a growing number of applications, particularly for machine learning. Investigators often collaborate with external entities and request that patient data be shared with organizations outside of UCLA Health. We report an approach to governing the release of patient data to third-party organizations.

Methods

The Figure shows an overview of our process. Staff work with submitting investigators to ensure that complete information is elicited for each release. Next, a Data Release Subcommittee, consisting of CTSI and UCLA Health representatives, reviews the request’s potential risk, including data security, the use of sensitive data such as drug use, and consistency with any IRB approvals. Next, UCLA’s Technology Development Group develops a contract with the third party, which must meet 16 mandated terms, the first of which is “Purpose must advance public benefit.” The request and the terms are then reviewed by the Health Data Oversight Committee, which consists of 12 members including the CIO, Chief Counsel, and CTSI Director. Finally, the agreement is approved or rejected by the Vice Chancellor for the Health Sciences, who oversees both the David Geffen School of Medicine and UCLA Health.

Results

In the first 18 months (9/2019 to 2/2021), investigators submitted 109 requests for external data releases. The rate of submittal has varied from 2 to 12 per month, with no distinct uptrend. Of the requests submitted, 33 were withdrawn by the investigator, 20 are actively in process and 29 are suspended due to non-response from investigators. For 27 requests, the review process has been completed by both committees and the Vice Chancellor, with 23 requests (85%) having been approved and 4 (15%) not approved. The poster will review examples in specific categories.

Conclusion

Our review process, which involves a risk assessment, followed by a value assessment with mandated contract terms, has resulted in 27 well-justified data releases over 18 months. Only 15% of requests completing the review process were not approved, but almost half of requests were withdrawn or suspended, suggesting that the process often leads investigators to decide on their own that the value-to-risk ratio is not sufficient to justify their request.
Supporting EHR-based Cohort Discovery Through User-centered Design: Results of an Early Formative Usability Study

Natalie C. Benda, PhD1, Pascal S. Brandt, MSc2, Jessica S. Ancker, MPH, PhD1, Jennifer A. Pacheco, MS3, Prakash Adekkanattu, PhD4, Guoqian Jiang, MD PhD4, Jyotishman Pathak, PhD1, Luke V. Rasmussen, MS3

1Weill Cornell Medicine, New York City, NY1; 2University of Washington, Seattle, WA; 3Northwestern University, Chicago, IL; 4Mayo Clinic, Rochester, MN

Introduction
Cohort discovery is a critical step in clinical and biomedical research studies. Computational electronic health record (EHR)-based phenotyping provides a means for cohort identification that limits the need for exhaustive chart review. Creating portable phenotypes that may be implemented and validated in different institutions presents a challenge in this space. One aim of the ongoing Phenotype Execution and Modeling Architecture (PhEMA) initiative is to develop an authoring tool for creating and executing portable phenotypes.1 To create a useful, usable tool, we have taken a user-centered design approach for system development. Here, we present initial results from a formative usability study to explicate user needs for supporting authoring and execution of phenotypes across institutions.

Methods
We first created a wireframe interface of our phenotype “workbench” through iterative task analyses and design cycles with the PhEMA team that has expertise in phenotyping and user-centered design (examples in Figure 1). The wireframes were pre-populated with sample phenotype algorithms, and we completed formative usability assessment interviews with representative end users who had experience analyzing EHR data. Participants granted written informed consent and were reimbursed for their time. Each participant was assigned the task of executing two phenotype algorithms without any formal instruction and as little help as possible. After working through the execution, a brief semi-structured interview was conducted to elicit additional suggestions and feedback. The Weill Cornell IRB approved this project.

Results
Five expert phenotype author participants found the textual summary generally useful, and noted that the graphical phenotype summary may be helpful for data requesters. Participants also expressed the utility of the results summary view. However, participants had design critiques for each interface component. Specifically, participants wanted the ability to link to terminologies presented in the textual phenotype summary and a means for drilling down to demographics in the results summary. The phenotypes in the study came “pre-loaded”, and participants further described that they would want the ability to load and edit additional phenotypes.

Discussion and Conclusion
The results of our formative usability assessments directed our efforts to implement two new features in our phenotype workbench 1) connection of the textual summary to standardized terminologies through the Value Set Authority Center and 2) the ability to directly load and edit phenotypes from the Phenotype Knowledge Base. Future efforts will involve cognitive task analytic approaches to gain a deeper understanding of phenotyping activities to provide enhanced visualization for phenotypes and their outputs for authors and their collaborators.

References
Colorectal Polyp Extraction System (CoPEx): A Natural Language Processing Pipeline to Extract Colorectal Polyp Characteristics from Pathology Reports

Ryen Benson1, Candace Winterton2, Maci Winn1,2, Mei Liu PhD, MS3, Noor Abu-el-rub PhD, MS3, John Hurdle MD, PhD4, Mike Conway PhD4, Andrew Gawron MD, PhD1,4, Sheetal Hardikar MBBS, PhD, MPH1,2

1University of Utah, Salt Lake City, UT, USA; 2Huntsman Cancer Institute, Salt Lake City, UT, USA; 3University of Kansas Medical Center, Kansas City, KS, USA; 4Salt Lake City VA Specialty Care Center of Innovation, University of Utah, Salt Lake City, UT, USA

Problem to be Addressed
Although several risk factors have been associated with colorectal adenomas, precursors to colorectal cancer, risk factors for less-studied polyp subtypes, including the more aggressive serrated polyps, are unknown (1,2). Identifying risk factors for these polyp subtypes is critical to developing evidence-based personalized screening strategies. However, studying the risk factors for these polyps is often cost- and resource intensive as polyp features are embedded within unstructured pathology reports. Consequently, Natural Language Processing (NLP) could serve as an effective tool to extract large amounts of unstructured data from pathology reports and enable construction of epidemiologic datasets to study risk factors for colorectal polyps.

Methods
Using pathology reports (n=12,515) from the University of Utah Enterprise Data Warehouse, we developed a rule-based NLP pipeline to extract colorectal polyp diagnoses and related features (number, site, shape, and size of colorectal polyps) for all colonoscopies at the Gastroenterology clinic at the University of Utah from 2013-2018. To develop and refine our annotation scheme and determine interrater agreement, two authors manually annotated 200 pathology reports for histopathologic features of colorectal polyps. Once a strong level of agreement was achieved and maintained (i.e., Cohen’s κ > 0.8), the same authors annotated a validation set of 150 pathology reports to serve as a reference standard for calculating performance metrics of our pipeline.

Results
We extracted the histopathologic features of 11,269 polyps across 6,323 patients with a colorectal polyp diagnosis (2,367 individuals were polyp-free) using our NLP pipeline. On average, patients had 1.7 colorectal polyps resected with an average largest polyp size of 0.54 cm. Our pipeline extracted 6,688 adenomas (98.2%, 1.7%, 0.03%, specified as a tubular, tubulovillous, or villous pathology, respectively), 1,208 serrated polyps, and 3,104 hyperplastic polyps. Of the adenomas, 5.6% were classified as advanced adenomas (i.e., polyps > 1 cm. in size or containing a villous or tubulovillous pathology), and 1.8% of polyps displayed dysplasia. Overall, our NLP pipeline achieved a precision of 98%, a recall of 91%, and an F1-score of 94% in extracting all colorectal polyp variables of interest.

Conclusions
We developed an NLP pipeline that accurately extracted the diagnosis of 6,688 adenomas, 1,208 serrated polyps, and 3,104 hyperplastic polyps and their clinically relevant features from unstructured pathology reports from 6,323 individuals undergoing colonoscopy at the University of Utah Gastroenterology clinic between 2013-2018. We have successfully demonstrated the utility of NLP techniques to create unique datasets for the epidemiologic research of colorectal polyp risk factors.

Acknowledgments
This work was supported by NCI K07 CA222060 (S. Hardikar) from the National Cancer Institute, T15LM007124 (R. Benson) from the National Library of Medicine, and institutional funding from the University of Utah.

References
Parent and Paraprofessional Use of Commercial Augmentative and Alternative Communication Devices with Non-verbal Autistic Young People

Ariana Bernstein, MS, Jina Huh-Yoo, MHCI, PhD, Andrea Forte, MS, PhD
Drexel University, Philadelphia, PA, USA

Introduction
In a report published in 2020, autism spectrum disorder affects 1 in 54 children as of 2016, a move from 1 in 60 in 2014 [1]. Approximately 30 percent of children with autism are minimally verbal [2], meaning they communicate using little or no spoken language. Novel technologies have introduced new communicative possibilities for minimally verbal individuals and the people close to them. We conducted semi-structured interviews to explore the experiences of people who use commercially available communication technologies to interact with minimally verbal autistic children. Our goals are to understand the experiences of teachers and parents who use these technologies, to learn from them about the experiences of their minimally verbal students and children, and to derive insights about future design and use of aided augmented and alternative communication (AAC) devices and software.

Methods
We conducted six semi-structured interviews recruited using convenience sampling to encourage parents and speech-pathologists of non-verbal individuals to raise relevant issues that we may not have anticipated and to allow for storytelling and rich descriptions of their general technology use and retrospective reflections on interactions with others using AAC devices. The team iteratively coded transcribed interviews beginning with open codes grounded in data, and gradually grouping and thematizing the codes to generate the findings. We received an IRB approval for in-person and remote interviews. Participants included three professionals who have and currently work with non-verbal individuals as aides and speech-pathologists (referred to as ‘‘aides’’ or ‘‘paraprofessionals’’) and three parents of non-verbal children. The interviewer asked for examples of people who stood out in participants’ minds and asked how communication happened with those individuals and probed the details about any technologies that were used.

Results
Common dedicated devices among the participants included: Accent Program with Unity, Bigmack, LAMP Words For Life, Proloquo2Go. We found four major themes regarding participants’ experiences with AAC and a methodological insight. (1) Customization: It was critical to choose what devices and strategies work for non-verbal individuals, as one participant explained: “We say their voice, their choice.” (2) Maintenance: When a dedicated device breaks, it needs to be sent back to the company and can take weeks to be returned to the individual. The severity of disruption associated with inconsistent access was detrimental to user experience. (3) Two-way Communication: Caretakers reflected that they, too, not just non-verbal individuals, need to learn how to communicate effectively with the devices. There was a lack of resources available for caretakers to learn how to use the device(s). (4) Usability: Low frustration tolerance is common among people on the spectrum and can lead to distress, which makes choice overload an egregious design problem that can erect critical barriers to use. Usability problems such as labeling issues and communication symbol sets contributed to this frustration. Lastly, we learned that the participants who work with non-verbal individuals closely in a 1:1 relationship and their concerted efforts to achieve intersubjective understanding with learners linked their experiences in a way that is uniquely powerful among informants in qualitative research.

Conclusion
In addition to the experiences we identified regarding commercial AAC devices, we discovered a methodological insight of how caretakers’ reports of their own experiences and concerns were entangled with those of the young people they taught and cared for. We elucidate the role of what we call intersubjective entanglement in the reported experiences of people who serve as caretakers and catalysts for the learning of others. This concept expands on the idea of key informants as individuals with access to a wide range of cultural roles and representations by virtue of their social relationships or position within an organization. We suggest that intersubjective entanglement is a useful concept to explain the role of key informants in interview-based work that seeks to extrapolate understanding or design implications for supporting a vulnerable population alongside data collection from vulnerable populations themselves.

References
Understand the Role of Health Literacy in Relation to Social Determinants of Health: A Systematic Review

Shwetha S. Bindhu1, Anunita Nattam1, Catherine T. Xu1, Tiffany Grant, PhD1, Hexuan Liu, PhD1, Danny T.Y. Wu, PhD, MSI1
1University of Cincinnati, Cincinnati, OH;

Introduction

In recent years, clinical note sharing has expanded patients’ abilities to understand and control their own health, especially for patients from vulnerable populations. Addressing low levels of health literacy (HL) is an important factor in improving the benefits of clinical note sharing, as high HL helps improve shared decision-making1. While causes of low HL have been extensively researched, the nature of health literacy as a modifiable social determinant of health (SDoH) is yet to be thoroughly explored. This systematic literature review aims to understand the nature of the modifiable relationship between HL and SDoH and the feasibility of developing interventions to address HL.

Method

The study systematically reviewed literature following PRISMA guidelines2. Literature in PubMed and Scopus databases were searched according to the eligibility criteria. The search criteria used were “health literacy” AND “social determinants of health.” Two researchers separately screened the papers by title and abstract. Disagreements were resolved by a third researcher after the initial screening. The process was repeated for full article eligibility. Papers that were included for the full article review were assessed for quality using AHRQ-informed assessments. After excluding articles with a “poor” quality rating, a team of four researchers extracted information from the remaining articles based on PRISMA guidelines and the articles’ SDoH and HL of focus2.

Results

A total of 146 articles from PubMed and 243 articles from Scopus were found through the search criteria, and 49 articles will be assessed in the narrative synthesis. Figure 1 shows the totals after each step of screening, eligibility, and quality assessment were conducted. Figure 2 shows the study design totals after the information extraction was completed.

<table>
<thead>
<tr>
<th>Study Design</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observational Study</td>
<td></td>
</tr>
<tr>
<td>Cross-sectional</td>
<td>43</td>
</tr>
<tr>
<td>Non-comparative</td>
<td>1</td>
</tr>
<tr>
<td>Qualitative</td>
<td>2</td>
</tr>
<tr>
<td>Random Controlled Trial</td>
<td>1</td>
</tr>
<tr>
<td>Literature Review</td>
<td></td>
</tr>
<tr>
<td>Integrative</td>
<td>1</td>
</tr>
<tr>
<td>Scoping</td>
<td>1</td>
</tr>
<tr>
<td>Systematic</td>
<td>1</td>
</tr>
</tbody>
</table>

Discussion

Current progress on the systematic literature review indicates that there has been substantial research into the multidimensional nature of the relationship between HL and SDoH and that significant consideration has been taken on the implications for intervention procedures. We will continue to conduct the narrative synthesis to summarize the large-scale impacts and relationships between these multiple factors and in improving clinical note sharing for patients with low HL.

References

Case study in the development of a framework for quality and reproducibility in inner-sourced packages and self-service analytic dashboards to accelerate common study types

James A. Black, PhD\textsuperscript{1}, Nayan Chaudhary, MSc\textsuperscript{2}, Adam J. Fory, PhD\textsuperscript{3}, Sriraman Madhavan, MSc\textsuperscript{3}, Matthew Secrest, MSc\textsuperscript{3}, Kamil Wais, PhD\textsuperscript{2}  
\textsuperscript{1}PD Data Science, Roche, Switzerland; \textsuperscript{2}7N Consulting, Poland; \textsuperscript{3}PD Data Science, Genentech, United States. All authors contributed equally.

Introduction and Principles

Reproducibility of a research question and reproducibility of individual studies are hampered in studies due to the diversity in the code that sits between a statistical analysis plan and a result. We propose a set principles that can be immediately implemented to boost efficiency.

Code has value: The use of ad hoc code within a study should be minimised, and a culture promoted of collaboration on pan-study code in documented packages for reuse. Unit tests are ideally written at function creation, and reviewed and expanded with new use cases.

Democratise access to analytics: Aided by the consolidation of code into packages, less technical users should be given access to well documented packages to promote guided analyses, as well as dashboards considered for fully self-service interaction.

Be verbose and specific on input cohorts: Publish cohort derivation code as user readable markdown vignettes with descriptive statistics and assumption checks.

Assertively limit user inputs in dashboards: Developers should ensure that variables exposed for manipulation in the app can be varied without compromising validation.

Separate logic from the user interface: As shown in , scientific logic should be separated from visualisation code. This loosely coupled approach improves the robustness of the system by making it easier to isolate, test and document any logic applied to the data.

Case study

By applying these principles to a common study data type, the estimation of duration of treatment in real world settings using routinely collected data, we were able to accelerate time to deliver study results from months to days, while improving robustness and standardisation via well documented and tested code.

In this case study we wrapped repeated study code into a robust and unit tested R package (Code has value), which contained all scientific business logic (Separate logic from the user interface). When then created a de-coupled interactive user interface via Shiny (Democratise access to analytics). Within Shiny we controlled inputs, to ensure the underlying R package was used within a valid scope (Assertively limit user inputs in dashboards). In addition to logic required for the final app, the R package contained vignettes that defined and provided documentation metadata for versioned cohorts. These versioned cohorts, and the accompanying metadata, were then ingested by the app (Be verbose and specific on input cohorts).

Discussion

These principles provided a mechanism to improve the reliability of our code, and had an important secondary effect of providing a clear framework to communicate with stakeholders and between data scientists steps taken to improve quality and reproducibility of internal study dashboards.
Challenges in Designing A Notification System for Electronic Medical Records: An Exploration of User Expectations

Katherine Blondon, MD, PhD
University Hospitals of Geneva, Geneva, Switzerland

Introduction
Information overload can occur in electronic medical records (EMRs). A notification system can help draw attention to the arrival of new data in a chart (notifications) or to abnormal or critical values (alerts). The aim of this paper was to explore what clinicians expect to be notified about when managing their patients and how they expect to be informed. We hypothesized an association between the perceived importance of a result and the alert mode.

Method
We recruited 9 doctors (7 supervisors, 2 residents) from different medical specialties of our hospital to participate in a focus group. Participants were asked to individually estimate the importance of incoming information in 9 vignettes, and to report how they wanted to receive this information (smartphone message, patient’s chart or multi-patient view on the computer). They discussed their answers and explored approaches to avoid alert fatigue. Vignette responses and notes during the session were analyzed qualitatively, comparing and contrasting findings to find common themes and design implications.

Results
Although participants perceived the importance of incoming information similarly, and expected similar communication modes, there was no clear consensus. Context and acuity of response were used to interpret the new results. Low urgency was more often associated with EMR alerts and high urgency with smartphone alerts. For outpatients, doctors underlined the need for alerts because test results are often not seen until the patient’s following appointment. Participants also suggested notifications for emergency department patients’ tests to help decrease consultation times. Participants agreed that targeting the patient’s doctor was a good way to avoid alert fatigue, but recognized the difficulty to reliably identify the proper target for each patient. Another approach to avoid alert fatigue was to be able to adjust alert thresholds for a division and/or according to individual preferences.

Design implications from the focus groups results are reported below (Table 1).

Table 1. Design implications for a notification system with a default setting and possibility of customization

<table>
<thead>
<tr>
<th>Perceived importance</th>
<th>Delay for response</th>
<th>EMR notification / alert mode</th>
<th>Smartphone alert mode</th>
<th>Customizable by user*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>In a patient’s chart</td>
<td>Multi-patient view</td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>None</td>
<td>Contextualized</td>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td>Important</td>
<td>None</td>
<td>Not interruptive</td>
<td>Not interruptive</td>
<td>Yes</td>
</tr>
<tr>
<td>Important</td>
<td>Urgent</td>
<td>Interruptive</td>
<td>Interruptive</td>
<td>Text</td>
</tr>
<tr>
<td>Critical</td>
<td>None</td>
<td>Interruptive</td>
<td>Interruptive</td>
<td>Text</td>
</tr>
<tr>
<td>Critical</td>
<td>Urgent</td>
<td>Interruptive</td>
<td>Text with snooze</td>
<td>No</td>
</tr>
</tbody>
</table>

Conclusion
To address user expectations, alert systems need a default setting that allows some customization (by division, and by personal preferences) for non-critical events. Targeting alert receivers to help avoid fatigue requires being able to reliably identify the right person.

References
Comparing Performance of Hip-worn and Wrist-worn Research-Grade Activity Monitors for Daily Activities

Katrina Boles, MS¹, Malaika R. Gallimore, RN, MPH¹, Chuka Emezue, PhD, MPH, MPA¹, Chelsea Howland, MSN¹, Amy Grimsley, MSN, RN, CCRN-K¹, Jo-Ana D. Chase, PhD¹, Allison B. Anbari, PhD, RN¹, LeeAnne B. Sherwin, PhD, MS, FNP-BC¹, Blaine Reeder, PhD¹

¹University of Missouri, Columbia, MO

Abstract
Increasingly available consumer-grade wearable devices represent great opportunities for their inclusion in research. Measuring the usability, function, and accuracy of wrist-worn smart watches helps inform best choices for device selection for participant studies. Researchers often test consumer-grade devices against a “gold standard” device such as the hip-worn ActiGraph wGT3X-BT (GT3X). The ActiGraph GT9X-Link (GT9X) is a newer research-grade device designed for wear on the wrist. However, there are known differences in the accuracy of wearable devices based on wear location. Therefore, it cannot be automatically assumed that data from hip- and wrist-worn research-grade devices are comparable for measuring activity. We compared the hip-worn GT3X to wrist-worn GT9X in two different week-long test sessions with adult testers (session 1: n=4 participants; session 2: n=5 participants) for daily free-living activities, segmented on high versus low activity, using the GT3X as a gold standard.

Introduction
In the Precision Smart Technologies and Applications for Rapid Translation (START) laboratory, we use a step-wise methodology to evaluate consumer-grade devices for function and usability. We capture activity data with multiple devices worn concurrently and make performance comparisons with the GT3X as the accepted standard. We use these early comparisons to identify which devices to test activity levels with more evaluators in the next phase. ActiGraph devices are the most used accelerometers in physical activity research studies and are accepted as the gold standard. The purpose of this study is to compare daily activity step counts from the hip-worn GT3X and the wrist-worn GT9X.

Methods
Evaluation data were collected during two one-week sessions. In the first test session, four evaluators concurrently wore the GT3X on the dominant hip and the GT9X on the non-dominant wrist, set to the default sampling frequency of 30 Hz. In the second session, five evaluators concurrently wore both devices in the same locations. Evaluators logged two days of daily activities in each session using an open write-in method for activity description, start and stop times. Using a 60-second epoch for each device, we calculated the sums of total steps for the timeframe of each activity tracked. We identified key terms in the descriptions for high activity - such as walk, exercise, and move - versus low activity - including sit, nap, sleep, and drive. Ambiguous activities were left unidentified. We calculated the absolute percent error using GT3X as the accepted and GT9X as the experimental values for each activity:%

\[
\text{% error} = \left| \frac{\text{accepted} - \text{experimental}}{\text{accepted}} \right| \times 100.
\]

Results
A scatterplot of step count comparisons shows the GT9X captured more steps per GT3X steps (Figure 1). It is difficult to tell from the visual whether high or low activity levels account for larger differences in step counts. The median absolute percent error for low activity (n=107) was significantly higher at 195.2% versus 82.8% for high activity (n=51). The wrist-worn GT9X identified movement as steps more frequently during times of low activity.

Discussion
We identified that the GT9X has a higher error rate for step tracking during low-level activity. Therefore, the ActiGraph GT9X-Link is not a substitute for the hip-worn ActiGraph wGT3X-BT as a gold standard research device. Future studies could compare the GT9X and GT3X while worn in the same place (either hip or wrist) though this endeavor may be of limited utility since the GT9X was designed as a smart watch device.

References
Quantifying the Period Prevalence of Autoimmune Diseases Across the United States

Gagana Borra¹, Joshua M. Landman, MS²,³, Randi Foraker, PhD, MA, FAHA²
¹College of Arts and Sciences at Washington University, St Louis, MO, USA; ²Institute for Informatics at Washington University School of Medicine, St. Louis, MO, USA, ³Division of Computational and Data Sciences at Washington University, St. Louis, MO, USA

Introduction

Millions of individuals across the United States (US) have an autoimmune disease (AD) and it is difficult to diagnose an AD as there are various symptoms. We hypothesize that we can aid patients in obtaining a quicker diagnosis by quantifying the prevalence of ADs; allowing physicians to make more efficient, accurate diagnoses. To establish estimates of the baseline prevalence of ADs in the US, we conducted a literature review (LR) to characterize the prevalence of over 160 ADs. From these baseline rates, we identified the top ten most prevalent ADs in the US and used current electronic health record (EHR) data to calculate prevalence estimates for ADs.

Methods

A list of ADs¹ was compared with a table of ADs compiled by the LR and via consultations with physician experts at our institution. We added 57 diseases to the list, resulting in a total of 166 diseases, of which 109 of 166 diseases were verified as ADs in our healthcare system. Search criteria for ADs since 2010 included “disease name” and “prevalence rate” (PR). Based on the LR, the top ten most prevalent ADs were identified. We extracted EHR data from our health system, using International Classification of Diseases Version 10 (ICD-10) codes and capturing cases diagnosed from 2010 to 2019 inclusive (n=6,494). We estimated the period PR as cases/100,000 persons. Using R statistical software², we age-standardized these rates to the US population and described the demographic distributions for all 109 ADs in our system’s EHR data during this timeframe (n=19,336).

Results

Rheumatoid arthritis had the greatest number of cases (n=2,925). Its PR in our health system was 170.5 (US 163.4). Lichen sclerosus had the fewest number of cases (n=71). Its PR in our health system was 4.1 (US it was 3.9). Patients with polymyalgia rheumatica had the oldest average age (80 years). Patients with urticaria had the highest variation in age (+/- 26 years). The top ten prevalent conditions were most likely to be diagnosed in women than men. The PRs for the top ten ADs in the US are seen in Figure 1. The blue bar shows age-standardized US rates. The red bar shows the published PR. *Psoriasis exceeds 1000 with a PR of 2328.

Figure 1. Top Ten Prevalent Autoimmune Diseases in the US

Conclusion

Quantifying the prevalence of ADs from 2010 to 2019 can serve as a reference point for future studies. By age-standardizing our locally estimated rates to that of the US population, we can observe how the PRs of these conditions may vary. Future research can update these rates, quantify the PRs of rare ADs which were missing from our analysis, and compare PRs across healthcare systems to assess the robustness of our prevalence estimates.

References

A comprehensive platform for induced pluripotent stem cell research data

Kirill Borziak, PhD¹, Irena Parvanova, PhD¹, Joseph Finkelstein, MD PhD¹
¹Icahn School of Medicine at Mount Sinai, New York, New York, USA

Introduction

Stem cell therapies, particularly those involving induced pluripotent stem cells (iPSCs), are a growing new therapeutic field for treating, repairing, or replacing patient cells, with the aim of addressing physical damage or disease. Our previous work examined how the voluminous amount of data generated by stem cell research is maintained across the many publically-available stem cell databases¹. However, the diverse forms of data resulting from stem cell research are not consolidated, stored, and available for access by researchers in a centralized and harmonized manner. Based on this necessity to homogenously organize and visualize stem cell data, we developed the Regenerative Medicine Data Repository (ReMeDy). The ReMeDy platform provides a unique solution by allowing the systematic aggregation of data by using a multi-modal common data elements (CDE) framework. The potential users of ReMeDy include clinical and pre-clinical researchers interested in understanding the similarities and trends in the derivation of iPSC products across different applications. The feasibility of ReMeDy has been tested using over 50 diverse, published iPSC projects. It is publically accessible at https://remedy.mssm.edu/.

ReMeDy platform and multi-modal CDE framework

ReMeDy is designed to be a user-friendly database, providing comprehensive and detailed information on iPSC projects. It is an implementation of the Signature Commons², available at https://github.com/MaayanLab/signature-commons. ReMeDy uses the data storage framework of Signature Commons, which is designed to store and search diverse metadata in an agile and flexible manner, using PostgreSQL for data storage. To further improve the usability of Signature Commons, an automated upload interface was developed using ReactJS and Spring Boot, which allows for public data upload and aims to promote crowdsourced uploading of published iPSC data into ReMeDy. To ensure strict quality control of the ingested data, all CDE values are submitted for validation prior to upload, using validator schemas against existing ontologies. The landing page provides easy access to the available functionalities of the platform: search, visualization, and API. Filtering schemas implemented in the search functionality allow users to incrementally refine their search query and provide statistical information on the distribution of CDE values. The API functionality allows researchers to download the data stored within ReMeDy as structured JSON format data.

To facilitate data collection and promote standardized data organization within ReMeDy, we have developed a multi-modal CDE framework that describes the full range of data generated by iPSC studies. Our framework aims to flexibly and comprehensively organize iPSC data, integrating iPSC and stem cell product characteristics, with patient and animal model information, omics data, and project findings and outcomes, to create a comprehensive framework of stem cell research data. The framework is organized into 5 modules: Project, Research System (patient and animal model CDEs), Manufacturing/Production (iPSC derivation and characteristics CDEs), In-depth Product Characterization (omics CDEs) and Outcomes/Findings.

The feasibility of the ReMeDy platform and the multi-modal CDE framework was tested by abstracting data from over 50 published iPSC projects manually by trained abstractors with experience in stem cell research. The abstraction focused on iPSC characteristics, patients or animal model and the research finding and outcomes. Our results indicate that ReMeDy is able to successfully accommodate in vitro, preclinical and clinical projects; data from both stem cell product and tissue engineered medical products; and mouse, rat, pig and rhesus monkey animal models.

Conclusion

ReMeDy provides a comprehensive, centralized and unified platform for data generated by iPSC research, as seen in our feasibility study. The improved access to aggregated CDE values and statistics has the potential to drive knowledge discovery. Future plans include automation of abstraction using natural language processing functionalities.

References

Challenges of comprehensive automatic coding of presenting complaints
Annie E Bowles, MS,1,2 Hannah Eyre, MS,1,2, Scott L DuVall, PhD1,2, Olga V Patterson, PhD1,2
1VA Salt Lake City Health Care System; 2University of Utah, Salt Lake City, UT, USA

Introduction
The Department of Veterans Affairs (VA) Informatics and Computing Infrastructure (VINCI) aims to improve access to VA clinical data and to facilitate their analysis. The VA research and operational efforts related to COVID-19 are supported through the VINCI COVID-19 Shared Data Resource (SDR)*. Emergency department (ED) presenting symptoms and chief complaints (CCs) have been identified as important in studying onset and progression of COVID-19 infection. An automatic coding algorithm that utilized the cTAKES dictionary annotator had been implemented to map the CCs as the primary reason for ED visit and admitting diagnosis to the UMLS 2020AA standard terminology. The output data is available in the COVID-19 SDR. While sufficiently accurate for coding the primary set of symptoms relevant for COVID-19 research, the complete set of collected mappings (referred to as CUIs) provided a valuable opportunity to evaluate mapping accuracy and clinical applicability of an existing algorithm on a large dataset.

Methods
As of March 1, 2021, the SDR has collected data for 1,242,560 patients and 601,622 ED visits, which resulted in 287,881 admissions. Processing 971,607 non-empty ED presenting symptoms, 77,592 ED discharge diagnoses, and 486,438 admit diagnoses, the mapping algorithm produced 1,682,270, 106,840, and 715,506 CUI mappings respectively. Validation was performed in three steps. First, we evaluated mapping accuracy against 1,000 manually mapped distinct CCs entries (100 most frequent and 900 randomly selected CCs). Second, we evaluated context of the mapped concepts on 400 CCs for negation, temporality, and certainty. Lastly, using the complete set of acquired mappings, we evaluated frequency of specific CUIs and semantic types to identify systematic errors of the mapping algorithm. ED presenting symptoms, ED discharge diagnoses, and inpatient admit diagnoses were evaluated separately.

Results and Discussion
Across the three evaluated datasets, the algorithm achieved 89.3-94.6% precision and 83.9-90.3% recall in identifying the correct CUIs. For the correctly mapped concepts, accuracy of context was measured as 80-89.7%. Misspellings, acronyms, concepts being mapped to multiple CUIs, and systematic mapping errors were the most frequent causes of errors. Due to the inherent ambiguity of short CCs statements, the mapping algorithm produces multiple CUIs that may or may not be related. For example, the concept “liver cancer” is double mapped to related CUIs: “Malignant neoplasm of liver” and “Liver and intrahepatic biliary tract carcinoma”. In contrast, the acronym “LBP” is mapped to both “lower back pain” and “low blood pressure”. Since the generic mapping algorithm attempts to map every possible concept, some CUIs may not carry clinical information such as “test” or “screen”. Varying concept granularity made accuracy comparison challenging. For example, did the algorithm fail if it mapped “acute” and “[disease]” as two CUIs, when a single CUI for “acute [disease]” exists? Context labeling evaluation showed concept variation that can only be differentiated through targeted processing. For example, “CUI=[test] context=[negated]” was identified for both statements test is negative and patient refused to be tested, even though the meanings of these statements are not identical. Manual review of most frequent CUIs on the complete dataset revealed a number of systematic mapping errors. ”ICD” as in ”ICD-10” was consistently mapped to an unrelated concept, and general use modifiers such as ”upper”, ”lower”, ”low”, ”high” were consistently mapped to CUIs related to specific diseases. Also, the vast majority of CUIs that had ”Nucleic Acid, Nucleoside, or Nucleotide” and ”Amino Acid, Peptide, or Protein” semantic types were erroneous. Systematic errors also affected recall. For example, a phrase ”failure to thrive” was never mapped to the CUI of the exactly identical UMLS concept. Once we identified a set of CUIs that were consistently erroneous, we implemented a post-processing step to remove them from the shared dataset, thus increasing its precision.

Conclusion
Ensuring data quality is a vital step in any study. By providing automatically extracted mappings and performing data cleaning on a common data set, VINCI has made a major step from facilitating access to existing data to providing previously inaccessible high quality variables to research groups.

Acknowledgement
This work was supported using VINCI resources and facilities, VA HSR RES 13-457.

References

An Open-Source Engine for Decision Support and Workflow Automation
Aziz A. Boxwala, MD, PhD, Mariano De Maio, Hank Wallace
Elimu Informatics, Inc., La Jolla, CA

Introduction
New health interoperability standards such as FHIR and CDS Hooks enable development of applications and services that can integrate with the electronic health record (EHR) and other clinical information systems. This integration allows healthcare organizations to extend the capabilities of the EHR to provide clinical decision support (CDS) and workflow tools to their staff and patients using apps and services that are internally developed or licensed from CDS vendors. While these new standards make it easier to integrate such applications, the development of the applications with complex logic and workflows still requires significant effort. Thus, in practice, few applications have been deployed for clinical use. To expedite the creation of interoperable applications, we developed a2d2, an open-source engine for implementing interoperable clinical automation services for CDS and workflow execution. The OpenCDS project provides open-source software for implementing CDS services.

Architecture and Implementation
An a2d2 server hosts and executes microservices for clinical automation. A microservice provides a well-defined interface and typically implements a single function as a module, e.g., a CDS Hooks service for computing pharmacogenomic interactions for a medication being prescribed. In a2d2, the work to be done by the microservice is specified in a process diagram in the standard Business Process Model and Notation. A process diagram is a set of tasks organized in a graph. The a2d2 software provides built-in tasks for querying and writing to FHIR APIs (of the EHR or other FHIR servers), calling non-FHIR APIs, parsing CDS Hooks requests and creating CDS Hooks cards, evaluating data against rules, sending text messages, transforming data into different models, and calling FHIR terminology services. The a2d2 software is implemented in Java. It uses JBoss’s jBPM Suite for the business process engine and JBoss Drools engine for rules evaluation. Process diagrams can be created using a web-based authoring workbench from JBoss. In our experience, the performance of the microservices is limited primarily by the response time of the FHIR APIs. For larger workloads, the a2d2 server can be horizontally scaled.

Results and Discussion
We have used a2d2 to create a variety of applications for different care settings that integrate with diverse EHR systems: (1) In an application for obtaining patient-reported outcomes (PROs), an a2d2 service was used for sending text messages to patients to complete the PROs at a physician-specified schedule. Another service received completed PROs, transformed them into documents, and wrote them into the patient’s chart in the EHR. (2) In an intraoperative CDS application integrated with the EHR, a2d2 services received asynchronous events such as intraoperative labs, and if needed generated alerts, and displayed them on the anesthesiologist’s monitor. Additional services provided drug dosing information and drug interaction alerts synchronously when medications were scanned for administration. (3) A SMART on FHIR application used in the ICU leverages an a2d2 service to prefill a form for the Sequential Organ Failure Assessment (SOFA) score with data from the EHR and another a2d2 service to calculate the SOFA score and perform risk stratification. (4) A pain management app used post-operatively in an acute care setting by surgeons, calls on a2d2 to compute from FHIR MedicationAdministration resources the dosage of opioids administered daily to a patient as morphine milligram equivalent. (5) An app for hypertension CDS uses a2d2 to obtain blood pressure data from a consumer health device and display the trends graphically. (6) A pharmacogenomic CDS application uses a2d2 to perform drug-gene interaction checking by querying specific regions of a patient’s whole genome sequence data, normalizing variants, computing genotypes, and creating FHIR DiagnosticReport resources.

Informaticians and analysts in our team have rapidly implemented complex automation logic for the services described above in a2d2. Externalizing this logic from the code of the applications has also allowed us to make and deploy updates to these automation services more frequently. We have found that a small number of built-in tasks described above have been sufficient to create many complex workflows due to the flexibility of the framework. We plan further enhancements to a2d2 to make development of services easier. We also intend to provide more support for standards such as supporting use of rules written in Clinical Quality Language and closer alignment with the work being done by the BPM+Health. The source code for a2d2 is available at https://github.com/elimuinformatics/a2d2.

References
Expanding the secondary use of prostate cancer real world data: Automated Classifiers for Clinical and Pathological Stage

Selen Bozkurt, PhD\(^1\), Tina Hernandez-Boussard, PhD\(^1\)

\(^1\) Department of Medicine, Stanford University, Stanford CA, USA;

Abstract

Cancer stage is essential to determine prognosis and treatment options, yet often recorded as unstructured text in EHRs. We developed an NLP solution to predict clinical and pathological TNM stages from EHRs from 5461 patients. We trained ML and rule-based algorithms to classify stage. We extracted TNM stages with high accuracy (average precision =90\% and recall = 86\%) and imputed 21\%-71\% of missing stage information in the institute cancer registry. Advanced NLP and ML technologies are necessary to capture staging information embedded in real-world data.

Introduction

Cancer stage is one of the most important parameters for cancer diagnoses, which guides treatment options and prognosis. However, this information is often not recorded in the electronic health record (EHR) as a structured field but instead found within the unstructured text and requires tedious manual effort to make this information available for secondary use. This presents a barrier to leveraging these real-world data sources for clinical research and clinical assertions. Therefore, adoption of advanced informatics methodologies such as machine learning (ML) and natural language processing (NLP) to unlock the vast information embedded in clinical free text and converting it into a usable structured format is essential to holistic use of the EHR.

Methods

We used data collected from a prostate cancer Clinical Data Warehouse (CDW), which links EHRs to the California Cancer Registry.\(^1\) As our primary guideline we used the most widely adopted cancer staging system for prostate cancer maintained by the American Joint Committee on Cancer (AJCC), which includes information on tumor, nodes, and metastasis.\(^2\) We developed two different NLP pipelines: 1) rule based and 2) semi-supervised machine learning. For each of T, N and M stages, we used the key-words based document-level vector representations of the text to train a classifier against the stage labels from the manually annotated set.

Results

There were 5461 prostate cancer patients in our initial cohort whose first line of treatment were completed in our hospital. Clinical stages of the patients were mostly T1 (44\%), N0 (61\%) and M0 (62\%). Likewise most of the patients’ pathological stages were T2 (51\%) and N0 (75\%). Since the number of patients whose pathological M stage is 1 were less than 10, these patients were excluded from the analysis. Overall, more than 20\% of staging information were missing for both clinical and pathological stages. Based on the distribution across all years, the highest number of missing cases were in 2018 (42\%). Manual chart review (‘gold’ standard) of the 1200 patients’ clinical and pathological reports were used for training and testing of the NLP models. For clinical T and N staging, rule based model outperformed the ML model with F1-scores over 0.71. However, for clinical M stage classification ML models reached better results than the rule based model with F1-score 0.98 for M0 and 0.88 for M1. For pathological T stage classification, both models achieved similar results with F1-scores over 0.85. ML model failed to classify N1 stages whereas rule based model reached F1-score of 0.88.

Conclusion

Cancer stage, a critical piece of information necessary to treat and prognosis newly diagnosed patients, is often not available in an easily obtainable format in real world data yet often missing. Our NLP pipeline to extract both clinical and pathological stages from clinical narratives and used this pipeline to augment staging documentation in our cancer registry.

References


Development of a multi-class deep-learning algorithm capable of diagnosing many histopathologic entities in digital pathology.

G. Thomas Brown¹, MD, PhD; Sudhir Sornapudi², PhD; Paul Fontelo³, MD, MPH
¹Artificial Intelligence Resource, NCI, NIH; ²Corteva Agriscience, Johnston, IA; ³National Library of Medicine, NIH

Introduction

The pathologist’s role in diagnosing cancer and guiding treatment is essential for patient care. As our understanding of cancer improves, the amount of knowledge required is increasing at a rapid pace, making the task of diagnosing tumors increasingly difficult, sometimes requiring expert consultation. Numerous papers have been published regarding digital pathology and machine learning¹⁻⁴, yet digital pathology tools remain primitive and not implemented on a wide scale. Although the practice of pathology covers a broad spectrum of diagnoses and organ systems, many reports have focused on limited specialties or diagnoses.

Methods

We have collected a wide variety of specimens to train a multi-class algorithm (VGG19) capable of assigning a wide range of diagnoses. We curated and digitized 2378 de-identified pathology slides (22 diagnoses). Pathologists digitally annotated regions of interest (ROI) indicating malignancies and patches (224 x 224 pixels) were extracted. Patches were also extracted from 112 benign slides. Three datasets were created at equivalent optical magnifications of 10X, 20X, and 40X. Each dataset was further split into 60% training, 20% for validation, and 20% for independent testing. Patches were only included in one arm of the dataset, i.e., a slide’s patches were not distributed across training, testing, or validation. We trained for over 200 epochs on a DGX-1 GPU server, and evaluated the model from the best epoch.

Results

The best results were generated with 20X magnification (Figure 1). The overall accuracy was 85% across the 22 classes, while the F1 score (micro) and Matthew Correlation Coefficient were 85% and 84%, respectively. When we trained at 40X magnification, the overall accuracy was 53%, while accuracy for training at 10X was 40%. At 20X, we achieved >90% accuracy for classifying clear cell renal carcinoma, mucinous carcinomas, granulosa cell tumors of the ovary, alveolar rhabdomyosarcomas, desmoid fibromatosis, pleomorphic sarcomas, urothelial carcinomas, and ganglioneuromas. Accuracy for angiosarcomas and malignant peripheral nerve sheath tumors was 30% and 25%, respectively. While 56% of lung squamous cell carcinomas were classified correctly, it erroneously labeled 26% of those cases as lung adenocarcinomas. Conversely, the algorithm correctly labeled lung adenocarcinomas 70% of the time, while misclassifying 14% of those cases as lung squamous cell carcinomas. The confusion matrices indicate 20X magnification yields the most accurate results.

Discussion

Current deep-learning algorithms generally focus on limited organ systems or diagnoses categories. We believe this is the first multiclass, deep-learning algorithm capable of diagnosing many histopathologic entities in general surgical pathology, with a resulting overall accuracy of 85% across the 22 classes. Although our algorithm performs well at 20X, it is unable to distinguish between different types of soft tissue tumors (e.g., sarcomas), which correlates with the difficulty human pathologists encounter, who often defer to sub-specialty trained pathologists. The algorithm also did not perform at a sufficient accuracy at differentiating between squamous cell and adenocarcinomas of the lung, corresponding to a major clinical error. The treatment regimens between the two are incompatible and could result in serious injury if the correct diagnosis is not made⁵. Finally, the algorithm misclassified benign nerve tissue into the general category of benign, which is not significantly incorrect clinically, but did count against its overall accuracy. Despite this, the overall accuracy was quite high. We plan to further expand the range of pathologic entities, which we believe will enhance its clinical utility, as well as in biomedical research. In countries where pathologists are in short supply, an algorithm such as this could aid oncologists and guide initial treatment, while waiting for confirmation. Likewise, practicing pathologists may be more confident in their diagnoses with a digital second opinion. Lastly, biomedical researchers could benefit from a “virtual pathologist”, as it can be difficult and/or expensive to solicit pathologist input.
How accurate is the Isabel diagnostic decision support system with patients presenting with acute symptoms of heart disease?

Katherine A. Brown, MSN, RN\textsuperscript{1}, Hamish S. F. Fraser, MBChB, MSc\textsuperscript{1}  
\textsuperscript{1}Brown University, Providence, RI, USA

Introduction

Diagnostic decision support systems have been developed to augment the diagnostic decision making process of clinicians over several decades and are intended to help improve patient safety by reducing diagnostic errors.\textsuperscript{1,2} Even as technology advances, algorithms have to date not been shown to outperform physicians, but when these CDS systems are used alongside the physician, they can improve the diagnostic accuracy of either one on their own. New electronic differential diagnosis (DDx) generators continue to be developed in recent years to augment clinician diagnosis, but more rigorous clinical evaluation based on real patient data and standardized evaluation methods is needed to best characterize their performance, accuracy, and ensure proper use by clinicians.

Methods

The purpose of this study was to measure the diagnostic performance of a well-established stand alone general medical diagnosis program designed for use by clinicians called Isabel Pro DDx Generator.\textsuperscript{3} This study used deidentified patient data from an earlier study that collected data on urgent admissions to cardiac services that obtained a differential diagnosis list from resident physicians and the specialized Heart Disease Program. The diagnostic performance of Isabel was tested from a cardiology perspective and compares its performance against the gold standard to that of resident physicians who collected and entered the data for urgent admissions.\textsuperscript{4,5} Clinical features from 57 patient cases (50\%) were entered into the Isabel Pro DDx Generator using a cardiologist validated protocol. After the checklist was generated, the top 20 diagnoses were recorded and compared to the final diagnoses (or “gold standard”) for each patient case from the original dataset. The performance metrics used for this study included Comprehensiveness (similar to sensitivity or recall), and Relevance (similar to Positive Predictive value or precision).\textsuperscript{6}

Results

Out of the 57 patient cases analyzed, 94.7\% had at least one match to the list of final diagnoses, 85.5\% had at least 2 matches, and 52.1\% contained 3 diagnoses matches by the Isabel generated suggestions. The mean Comprehensiveness and Relevance for the Isabel evaluation was reported for diagnoses generated at top 5, 10, and 20 thresholds. The Comprehensiveness measure is the proportion of correct diagnoses or matches to the total number of final diagnoses (similar to sensitivity or recall). The Comprehensiveness mean for all 57 patent cases was at 41.8\% for the top 5 output of diagnoses. It rose to 56.5\% for the top 10, and to 69.5\% for the top 20 threshold. Relevance is the proportion of matches to the total output list (or threshold maximums). The reported Relevance for all 57 patent cases was at 39.5\%, 28.7\%, and 23.1\% for the top 5, 10, and 20 thresholds respectively.

Discussion and Conclusion

The results show a majority of cases had at least one of the final diagnoses was present somewhere in the top 20 of the Isabel generated checklist of possible diagnoses. Comprehensiveness (top 10) matched those from the heart disease program and was higher than the physician performance from the previous study. This work along with further standardization of protocols, automated diagnosis matching strategies, and use of natural language processing can make strides toward large evaluation studies using real patient data extracted from Electronic Health Records.

Acknowledgements

This study did not receive external funding. The developer of the software did not have any role in the study.

References

Estimating Early Warning System Accuracy Prior to Implementation

Lusha Cao, PhD MS1, Gerald Shaeffer, MHI1, Meghan Galligan, MD1, Fuchiang Tsui, PhD1, Robert Grundmeier, MD1, Vinay Nadkarni MD1, Robert Sutton, MD MSCE1, Christopher Bonafide, MD MSCE 1, Naveen Muthu, MD1

1The Children’s Hospital of Philadelphia, Philadelphia, PA

Introduction

Critical deterioration events (CDEs) are life-threatening patient safety events in inpatient pediatric care that are challenging to recognize1. A pediatric early warning score (EWS) uses patient data such as vital signs to predict deterioration risk aim to help clinicians identify children at risk of deterioration1. While there has been a substantial amount of work to develop novel pediatric EWSs, less effort has been devoted to estimating the value of these scores prior to implementation in clinical care. Clinical care processes are already able to identify these patients (e.g. clinicians at our institution designate patients as “Watchers” when they are concerned), and in order to predict the value of a specific EWS2 for use by a freestanding intensive care rapid response team, we aimed to retrospectively estimate the accuracy of the EWS in addition to existing clinician recognition of at-risk “Watchers”.

Methods

Using data extracted from our institution's clinical data warehouse, we constructed a one-year dataset of inpatient encounters. We estimated EWSs at 4-hour time windows throughout each encounter by using the most recent relevant clinical observation and weighting according to the BedsidePEWS, a previously validated EWS2. This dataset was validated by manual chart review of 5% of CDEs by two authors (NM, MG). Sensitivity and positive predictive value (PPV) of the EWS were calculated based on the identification of CDEs in the 2-24 hours prior to the event.

Results

The cohort included 23,835 pediatric patients hospitalized for at least 4 hours in a medical or surgical ward, with 296 CDEs. High-risk patients identified by clinicians as “Watchers” was 40.9% sensitive for deterioration. When an EWS >=5 is added to clinician recognition (i.e. “Watchers”), 80.7% of CDEs were detected (Figure 1, last column). PPV of clinician “Watcher” identification was 16%, which was higher than either an EWS threshold ≥ 5 (3%) or ≥ 8 (8%). Estimating the predictive value of an “alarm” sent to a clinician every four hours for high-risk patients, we found that 14 Watcher alarms would fire for every one patient who deteriorated. Similarly, if EWS threshold was set at >= 8, then 14 alarms would fire and at >=5 then 33 alarms would fire for every one patient who actually deteriorated.

Figure 1. Proportion of critical deterioration events identified by EWS and/or “Watcher” labeling by clinicians.

Conclusion

Retrospective estimation of EWS accuracy suggests there is added value over clinician recognition, but the number needed to alarm is very high. This suggests that alternative approaches such as informing a proactive critical care outreach team may be a better use of an EWS risk score than sending an alarm to bedside teams. Such retrospective estimation of the value of prediction models should ideally be employed prior to EWS implementation into actual use.

References

Strategies to Engage Traditionally Marginalized Patients in Patient Portal

Heidi Carpenter1, Jacqueline Antoun1, Roman Gusdorf1, Austin Triana1, Elisa Friedman, MS2, S. Trent Rosenbloom, MD, MPH, FACMI, FAMIA2

1Vanderbilt University School of Medicine, Nashville, TN
2Vanderbilt University Medical Center, Nashville, TN

Description of the Problem: As patient access to electronic health records increases, older patients are often left behind due to less familiarity with computer use1. In addition, disadvantaged patient groups such as low-income, uninsured, non-English speaking, Black, and Hispanic patients are less likely to adopt the use of electronic patient portals2-3. My Health at Vanderbilt (MHAV) is a well-adopted electronic patient portal that has been in use at Vanderbilt University Medical Center since 20044. Patients can use MHAV to make appointments, access laboratory and imaging results and communicate with providers. Utilization of patient portals facilitates direct contact with the health care system and promotes transparency with reduction of errors; in addition, it increased patient connection, satisfaction, and treatment adherence4-5. To address barriers and disparities in MHAV access, a student-run initiative during the Martin Luther King Jr. day of service was created. Our aim was to target historically marginalized groups to improve MHAV adoption among diverse populations in preparation for COVID-19 vaccination opportunities.

Methods: Team members developed a database of all Vanderbilt University Medical Center patients ≥75 years old who did not have MHAV accounts. This patient list was organized by race and ethnicity to allow prioritization of patients from historically underserved populations. For this project, patients prioritized for enrollment were African American or Black race, Native Alaskan, American Indian or Hispanic ethnicity. Participating medical students contacted patients via telephone using a standard pre-approved script. The results of each encounter were categorized by patient contact status and MHAV status. Potential barriers were also coded, including access to hardware, technological literacy, waiting for assistance from children/caregiver, and language/hearing status barrier.

Results: During our student-wide day of service, 573 calls were placed. 86% were to Black or African American patients (n=492), 13% to Hispanic patients (n=76) and 1% to American Indian or Alaskan Native (n=5) patients. Calls answered in each group were 40% (n=196), 35% (n=27) and 75% (n=4) by race, respectively. In total, 3% of calls placed (n=17) resulted in successful sign-ups for MHAV. Of patients that were not successfully signed up (n=194), 18% (n=35) cited technological access or knowledge as the main barrier and 12% (n=24) took information about MHAV to sign up with a caretaker at a later time. Follow-up calls remained low yield.

Discussion of Results: This small pilot program identified that cold-calling older patients in historically marginalized groups has underwhelming efficacy for eliciting sign-ups to patient portal systems. Despite immense effort, most calls were unanswered. Those that did answer were not likely to sign up with lack of interest and lack of technological access as the two most common reasons.

Conclusion: Equal access to personalized health information and communication with providers is an important component of advancing health equity. However, cold-calling patients proved an ineffective method to engage older patients of traditionally marginalized groups in signing up for MHAV in preparation for COVID-19 vaccination efforts. Many cited access to technology and lack of technological literacy as the main barriers. Many planned to rely on others to access MHAV.

Attendee’s Take-away Tool: Cold-calling patients 75 years and older from marginalized groups is an ineffective strategy to elicit sign-ups for online patient portals in preparation for COVID-19 vaccination efforts.

References

1628
Sepsis Prediction Using Semi-Supervised and Transfer Learning

John R. Caskey, PhD1, Fereshteh S. Bashiri, PhD1, Anoop Mayampurath, PhD2, Nicole Dussault2, Jay Dumanian2, Sivasubramanium V Bhavani, MD3, Kyle A Carey, MPH2, Emily R Gilbert, MD4, Christopher J Winslow, MD5, Nirav S Shah, MD5, Dana P Edelson, MD, MS2, Majid Afshar, MD, MSCR1, Matthew M. Churpek, MD, MPH, PhD1

1University of Wisconsin, Madison, WI; 2University of Chicago, Chicago, IL; 3Emory University, Atlanta, GA; 4Loyola University, Chicago, IL; 5NorthShore University HealthSystem, Evanston, IL

Introduction
Sepsis is a life-threatening response to infection and is a leading cause of in-hospital mortality. Early treatment can improve outcomes, while avoidance of unnecessary antibiotics in non-infected patients can avoid complications. Electronic health record data can be used to identify infected patients, but these data contain discrepancies like antibiotic orders for uninfected patients. Manual chart review can determine infection status but is time-consuming for building large cohorts. Semi-supervised learning methods can take advantages of data without labels, while transfer learning can utilize data labeled with surrogate definitions to improve accuracy. In this study, we compare semi-supervised and deep learning to identify patients who are hospitalized due to an infection.

Methods
We conducted a retrospective cohort study at six hospitals in Illinois between 2006 and 2018. A random selection of 2,724 charts of patients who received either antibiotics or blood cultures were reviewed to determine infection on admission status (labeled), with 819 (30%) of these randomly selected as the test set. The remaining 430,241 admissions not selected remained unlabeled. Data from the first 24 hours were used to predict infection status, with characteristics, vital signs, and laboratory values used as features. XGBoost was utilized for the following semi-supervised methods: (1) self-supervised learning, where predicted probabilities were generated in the unlabeled data using a model fit to the labeled 70% training data, with highly confident predictions relabeled; and (2) cluster-then-label, where the labeled and unlabeled patients were clustered using k-means, then the unlabeled patients were relabeled based on the majority class of the labeled patients. An XGBoost model was also fit by assigning infection status to the unlabeled data based on whether the patient met the Sepsis-3 Task Force definition of suspected infection (antibiotics plus cultures). Next, we constructed combined convolutional neural network and long short-term network (CNN-LSTM) models that (1) used the same Sepsis-3 approach for labeling, and (2) performed transfer learning by fitting to the unlabeled data using the Sepsis-3 outcome and then fine-tuning in the 70% labeled training data. All models were compared to each other and to baseline XGBoost and CNN-LSTM models that only used the 70% labeled training data using the area under the receiver operating characteristic curve (AUC).

Results
The XGBoost baseline model had an AUC of 0.76, and none of the semi-supervised learning methods improved performance (Figure). The CNN-LSTM baseline model performed better (AUC 0.78) than the XGBoost baseline model (0.76) and XGBoost with Sepsis-3 labeling (AUC 0.74). The CNN-LSTM model fit using the Sepsis-3 outcome obtained the highest AUC (0.81), which did not improve when using transfer learning (AUC 0.81).

Conclusion
CNN-LSTM models more accurately identified infection than XGBoost. Semi-supervised methods did not improve performance, nor did transfer learning, with the CNN-LSTM model using Sepsis-3 labels having the highest AUC.
Automatic Detection of Surgical Site Infections Using EHR Data

Arjun Chakraborty, MS, Kevin Lybarger, PhD, Dustin Long, MD, Vikas N. O'Reilly-Shah, MD, Meliha Yetisgen, PhD
University of Washington, Seattle, WA, USA

Introduction

Surgical Site Infections (SSI) cost $3.5-10 billion and affect 160-300 thousand patients per year in the United States. SSI surveillance is usually done through labor intensive manual patient chart review. Automatic SSI detection has been proposed using machine learning with electronic health record (EHR) data. We explore the prediction of SSI in orthopaedic surgeries using a deep learning architecture that utilizes structured data and clinical text data from the EHR. A model which retrospectively predicts SSI could be used to assess the effectiveness of hospital infection control efforts or monitor the rate of surgical site infections at a hospital. The neural SSI predictor achieves high performance, 0.81 F1.

Methods

We use an existing dataset of 2,611 orthopaedic surgeries performed at the University of Washington Medical Center and Harborview Medical Center between 2008-2020. This dataset includes gold standard labels for SSI that were manually assigned to each surgery as part of the ACS National Surgical Quality Improvement Program (ACS NSQIP). The distribution of the SSI labels is imbalanced with 168 SSI-positive (6.4%) and 2,443 SSI-negative (93.6%) cases. We explore two feature sets: structured and text. The structured feature set includes nine laboratory fields from 7 days preceding to 90 days following each surgery. For the structured feature set, missing and outlier values are mean imputed. The text feature set includes all the clinical notes from 7 days preceding to 30 days following each surgery. Clinical text is encoded using term frequency-inverse document frequency (TF-IDF) unigram features. The vocabulary size is 128827. Samples in the dataset were randomly assigned into train, validation (for hyper-parameter tuning), and test sets with 3:1:1 a ratio.

We explore the prediction of SSI, as a binary prediction task, using two machine learning architectures: neural network (NN) and random forest (RF). RF is an apt baseline for this task, as several previous works which attempt the same task have used an RF classifier. For the structured feature set, the NN includes a single output layer (size=1). For the structured+text feature set, the NN includes three hidden layers (sizes=1000, 100, and 15 respectively) and an output layer (size=1). The NN configuration includes the Adam Optimizer, initial learning rate (0.001), number of epochs (10), RELU activation in the hidden layers, and sigmoid activation in the output layer. We experimented with up sampling positive cases and down sampling negative cases to account for class imbalance, but performance did not improve. We implement the RF model as a baseline for the NN architecture. The RF configuration includes the maximum depth (25 for structured, 40 for structured+text), minimum samples per split (4 for structured, 6 structured+text), # estimators (50 for structured, 200 for structured+text), and maximum features (all features).

Results and Conclusion

Table 1 presents the precision (P), recall (R), and F1-score (F1) of RF and NN models on the withheld test set for the structured and text feature set combinations. The NN outperforms the RF for both feature sets. Including the text features improved precision and recall for both models. This indicates that clinical text notes augment the information about SSI contained in structured data fields. The best performing model was the NN model with the structured and text feature sets. As future work, we will perform a detailed error analysis on this data set, investigate ways to improve the performance, and evaluate the generalizability of the proposed approach to other surgery types.

Acknowledgements

This work has been supported by NLM - 5T15LM007442-19 Biomedical and Health Informatics Training Program.

<table>
<thead>
<tr>
<th>Model</th>
<th>P</th>
<th>R</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>RF+structured</td>
<td>0.70</td>
<td>0.64</td>
<td>0.67</td>
</tr>
<tr>
<td>RF+structured+text</td>
<td>0.72</td>
<td>0.72</td>
<td>0.72</td>
</tr>
<tr>
<td>NN+structured</td>
<td>0.71</td>
<td>0.70</td>
<td>0.70</td>
</tr>
<tr>
<td>NN+structured+text</td>
<td>0.79</td>
<td>0.84</td>
<td>0.81</td>
</tr>
</tbody>
</table>

Table 1: SSI prediction performance
A Generalizable and Scalable Distributed Processing Architecture for Guideline-Based Clinical Decision Support

Justin Chambers, BS\(^1\), Tanya Podchiyska, MS\(^1\), Samson W. Tu, MS\(^{1,2}\), Amy Robinson PharmD\(^1\), Susana Martins MD MSc\(^1\), Michael Ashcraft MD\(^1\), Mary K. Goldstein, MD MS\(^{1,2}\)

\(^1\)VA Palo Alto Health Care System, Palo Alto, CA; \(^2\)Stanford University, Stanford, CA

Introduction

We developed the Medication Safety (MedSafe) clinical decision support (CDS) system to provide a production platform for integrated population-based clinical quality assessment and guideline-based CDS for individual patients in the domains of hypertension, type-2 diabetes mellitus, and chronic kidney disease. The MedSafe CDS system encodes clinical practice guideline (CPG) recommendations and drug information as knowledge bases (KBs). Applying an encoded guideline KB to individual patient data to generate patient-specific recommendations is a resource intensive task. In the MedSafe context where the clinician-facing performance measure dashboard may request a very large number of patients at once, a one-patient-at-a-time approach would not be able to finish in allowed time window. Therefore, we distributed the workload across multiple execution instances linked through a network engine and decoupled each execution instance into its own environment to avoid threaded conflicts within a process accessing resources. Furthermore, we added means to schedule jobs, accept requests, and report build statuses.

MedSafe Processing Architecture

We developed a component-based architecture that ensures component decoupling and reuse. The components are:

- Dispatcher/Listener service (Swarm API) – A generic client interface, which we implemented as a web server to dispatch processing requests and react to returned events. In response to a request, it partitions the requested workload and then distributes and balances it over the available resources. For recurring tasks, a job scheduler was integrated into the system to trigger maintenance and build operations.
- Resource manager (Hive) – A manager of the local health of drone workers that, over time, terminates drone workers that are not being utilized and refreshing instances when they are needed again. The distributed hives forms a computing cluster of multiple servers that communicate with the Swarm API.
- Worker client (Drone) – A component that encapsulates the execution environment for running any algorithm independently without external interference. In our case, it encapsulates the EON execution engine\(^1\), which loads the encoded-guideline KBs and applies them to the patient cases assigned to them. The drone worker receives processing requests from the dispatching host and communicates its status back during processing.
- Data source – A replaceable component that, in our implementation, uses the VA SQL Field Reporting Enclave (FRE) to accesses VA data and, when processing has completed, returns the resulting recommendations back to the SQL database.

While this architecture has been designed to support batch processing requests from a population-based performance measure dashboard, it can also respond to real-time requests as the client interface, upon receiving a request, may call the Swarm API to dynamically check for availability of a valid existing recommendation, and, if not available, call on the other components to process the single patient case.

This architecture was inspired by distributed computation systems such as Pixar’s Tractor\(^2\), a modern and robust solution for network rendering. The functionalities of this processing architecture are orthogonal and compatible with contemporary CDS standards. For example, the drone workers could be configured for the execution of Arden Syntax modules or Clinical Reasoning artifacts written with Clinical Quality Language. The Swarm Web Server can be adapted to field CDS Hooks requests and return CDS Hooks cards.

References


Views expressed are those of the authors and not necessarily of the Dept of Veterans Affairs.
emrKBQA: Creating a Clinical Knowledge-Base Question Answering Dataset

Rachita Chandra, MS¹,⁴, Preethi Raghavan, PhD¹,⁴, Jennifer J. Liang, MD²,⁴, Diwakar Mahajan, MS²,⁴, Peter Szolovits, PhD³,⁴

¹IBM Research, Cambridge, MA, United States; ²IBM TJ Watson Research Center, Yorktown Heights, NY; ³Computer Science and Artificial Intelligence Lab, Massachusetts Institute of Technology, Cambridge, MA, ⁴MIT-IBM Watson AI Lab

Introduction

The longitudinal, domain-specific nature of Electronic Health Records (EHRs) along with privacy concerns makes it difficult to develop large-scale annotated datasets for training machine learning models. This motivated the creation of emrQA¹, the first community-shared patient question-answering (QA) dataset, which was developed using a semi-automated process leveraging i2b2 annotations² to answer questions on unstructured clinical notes. Here, we present emrKBQA (complementary to emrQA), a dataset for answering questions on structured EHR data using MIMIC-III.

Methods

We generate emrKBQA by first collecting emrQA question templates and logical forms that were created from questions posed by physicians and subsequently normalized, e.g. “What is the dosage of insulin?” was converted to the question template “What is the dosage of [medication]?” These are then mapped to a corresponding logical form template. Second, question templates not answerable by MIMIC-III structured data are filtered out, and the entities and attributes in the remaining question and logical form templates are then mapped to the MIMIC-III schema. For example, since MIMIC-III does not capture relational information across tables, questions requiring such relations such as “Why did the patient have [treatment]?” are filtered out. Third, each question template is assigned one or more question types (YesNo, Temporal, Factual) depending on the kind of information it is seeking. Finally, question and logical form templates were populated with values from the MIMIC-III data tables, and the answer evidence extracted. The logic for these steps was implemented using SQL and shell scripts. Figure 1 shows an example of this process.

Results

Currently, emrKBQA consists of 940,713 question answer pairs over 100 patients, generated from 389 question templates and 52 question type-specific logical form templates. emrKBQA contains an average of 7.5 paraphrases per question type-specific logical form template (ranging from 1 to 55), where a paraphrase is defined as question templates sharing the same question type that map to the same logical form template. Of the generated question answer pairs, 90.9% are test results, 7.8% relate to medications, 1.2% to conditions, and the remaining to other topics (e.g. allergies, tobacco use).

Conclusion

In this work, we present emrKBQA, a dataset for QA on structured EHR data. emrKBQA is well suited for answering factoid questions since explicit values are well captured in tables, which may be difficult to infer from unstructured data. We believe emrKBQA is a pivotal contribution in the space, as together with emrQA, it provides the opportunity to explore differences between datasets derived from structured and unstructured clinical records, and to work towards a complete EHR QA system that considers data across both structured and unstructured data sources.

References


Figure 1. Process for generating emrKBQA
Towards Clinically Relevant Explanations for Type-2 Diabetes Risk Prediction with the Explanation Ontology

Shruthi Chari, MS, Prithwish Chakraborty, PhD, Oshani Seneviratne, PhD, Mohamed Ghalwash, PhD, Daniel M. Gruen, PhD, Daby Sow, PhD, Deborah L. McGuinness, PhD
1 RPI, Troy, NY; 2 Center for Computational Health, IBM Research, Yorktown Heights, NY

Clinicians who interact with clinical support systems often seek support beyond the displayed predictions - especially when seeing atypical patients or those who are not responding as expected. Depending on the context, clinicians summon different explanations (rationales) for reasoning, including contrasts, context, and evidence from the literature. Here, we present our vision to support user-centered explainability in a chronic disease comorbidities risk prediction setting where multiple data streams are typically used, leveraging a semantic resource we developed, the Explanation Ontology (EO).

In a type-2 diabetes (T2D) risk prediction use case, we are developing a pipeline to support a variety of explanations with different foci to explain the model predictions in a user-comprehensible manner. For risk prediction, we apply knowledge-infused Recurrent Neural Network models to assess the risk for comorbid complications within 365 days of each claim. Using post-hoc explainers, such as those from AIX 360 toolkit (https://aix360.mybluemix.net), we can generate explanations of such predictions at both model (e.g. global feature importance) and patient-level (e.g. personalized feature importance). In a proof-of-concept (POC) implementation, we contextualized risk prediction of a T2D comorbidity, Chronic Kidney Disease (CKD), through a question-answer approach to answer a pre-canned set of questions via feature importances and domain knowledge from the pharmacological chapter of the T2D guidelines. Building on this POC, we aim to enable user-centered explainability by leveraging the EO to present AI model components and relevant literature in a clinician-friendly manner (Fig. 1). The design allows clinicians to probe individual sections that highlight the different data streams for explanations. For instance, they might want to know “why a low risk for CKD, but a high risk for heart failure?” or probe the guidelines or cohort to understand “what was done for a similar patient case?” In part B of Fig. 1, we show an annotated example of an explanation generated by EO’s schema that elaborates on a risk prediction for a prototypical T2D patient. Such annotations can further help clinicians inspect explanations. We are currently extending the POC to support explanations to a more extensive set of questions for a T2D comorbidities use case.

References
AM2BERT: attention guided and regularized transformer-based multi-label classification model for COVID-19 literature curation

Qingyu Chen, PhD†, Jingcheng Du, PhD2, Alexis Allot, PhD1, Zhiyong Lu, PhD1†

1. National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health (NIH), Bethesda, MD, USA
2. School of Biomedical Informatics, UT Health, Houston, TX, USA
† To whom correspondence should be addressed: zhiyong.lu@nih.gov

Abstract

Every month, ~10,000 new articles are added to LitCovid, a literature database of COVID-19 related papers in PubMed; this rapid growth challenges the LitCovid curation pipeline. A primary bottleneck is annotating each article with up to eight possible topics. This study proposes AM2BERT, a transformer-based multi-label classification model. It achieves the highest overall performance (up to 8% higher macro-F1 score) and only requires ~15% of the inference time compared with the current best model.

Introduction

The number of articles in the literature related to COVID-19 is growing by about 10,000 articles per month (1). LitCovid (2, 3), a literature database of COVID-19 related papers in PubMed, has accumulated a total of more than 100,000 articles, with millions of accesses each month by users worldwide. LitCovid is updated daily, and this rapid growth significantly increases the burden of manual curation. In particular, annotating each article with up to eight possible topics, e.g., Treatment and Diagnosis, has been a primary bottleneck in the LitCovid curation pipeline.

Background

Innovative text mining tools have been developed to facilitate biomedical literature curation for over two decades (4). Topic annotation in LitCovid is a standard multi-label classification task, which aims to assign one or more labels to each article (5). To facilitate manual topic annotation, we previously employed the deep learning model Bidirectional Encoder Representations from Transformers (BERT) (6). We used one BERT model per topic, known as binary relevance BERT, and previously demonstrated this method achieved the best performance of the available models for LitCovid topic annotations (3); other studies have reported consistent results (7). However, this method has two primary limitations. First, by training each topic individually, the model ignores the correlation between topics, especially for topics that often co-occur, biasing the predictions and reducing generalization capability. Second, using eight models significantly increases the inference time, causing the LitCovid curation to require significant computational resources.

Method and results

Therefore, this paper proposes an Attention guided and regularized Multi-label classification model via Multi-task training and BERT as the backbone, which we refer to as AM2BERT. AM2BERT shares the BERT representations among the topics but uses a novel multi-channel attention mechanism to learn topic-specific representations. It also regularizes the representations via co-topic attention and auxiliary tasks. We compared AM2BERT to three baseline methods using two sets of evaluation metrics (label-based and example-based) commonly used for multi-label classification (8). It achieved the highest performance on the test set (2% higher macro F1-score and up to an 8% higher F1-score on a single topic than binary relevance BERT), and evaluating its generalization shows that it is much more robust to new articles (over 8% higher macro F1-score). Importantly, it requires only ~15% of the inference time needed for binary relevance BERT, significantly improving efficiency. AM2BERT has been employed in the LitCovid production system, making the curation more sustainable. We also make the datasets publicly available to the community via https://ftp.ncbi.nlm.nih.gov/pub/lu/LitCovid/topic_tagger/.

Localize the implementation of FDA list of confusable drug names with the application of Taiwan National Health Insurance database: a study of text similarity computation with real-world data

Ya-Lin Chen, PharmD1, Hsuan-Chia Yang, MSc, PhD1++, Der-Ming Liou, PhD1++
1Graduate Institute of Biomedical Informatics, College of Medical Science and Technology, Taipei Medical University, Taipei, Taiwan; ++: Yang, H.C. and Liou, D.M. contributed equally to this work

Introduction: Food and Drug Administration List of Established Drug Names Recommended to Use Tall Man Lettering (FDA list) has been widely implemented in Taiwan. However, a report from Taiwan government indicated that incorrect drug prescription due to name confusability is the main reason for patient harms. Drugs with similar names and frequently co-prescribed are at higher risk of contributing to medical errors due to cognitive confusion. We proposed a method to incorporate FDA list, text similarity computation and the real-world data of National Health Insurance database from Ministry of Health and Welfare of Taiwan, to optimize FDA list use in Taiwan by differentiating risk levels of cognitive confusion among drug pairs in the list.

Method: Text similarity of each pair of drugs in the FDA list was calculated by Python Levenshtein package. Drugs in all pairs were then labelled with WHO ATC codes and linked with co-occurrence counts and MMQ [1]. Co-occurrence is defined as the frequency of two drugs appearing in the same prescription; MMQ is defined as the ratio of joint probability of two medications in the database and the multiplication of the respective probabilities of two drugs. The two drugs whose respective probabilities are independent would have MMQ value of 1.

Results: After mapping drug pairs in the FDA list to WHO ATC codes, forty distinct pairs of drugs with administration routes as oral or parenteral were found. The median value of Levenshtein ratio among the pairs was 0.595. Fifteen drug pairs were further matched with co-occurrence and MMQ (Figure 1). Two thresholds were identified for risk categorization: the median value of co-occurrence and the value 1 of MMQ, represented as dotted lines in Figure 1. Four risk levels of cognitive confusion, Group A, B, C, and D, represent potential risk from high to low, based on the two decision thresholds. Drug pairs with red dots represent those with Levenshtein ratio more than 0.595, and the pairs with blue dots represent those with ratio less or equal to 0.595.

Conclusion: More than half of the drugs in FDA list are used in Taiwan, which highlights the importance of FDA list use. The optimization of the list use is valuable in tailoring international guidelines into local needs. This study developed a decision process for such purpose with risk levels of confusion. The study result can be further utilized for pharmacy storage layout adjustment or embedded in clinical decision support systems for prescription evaluation.

Reference

Figure 1. Drug pairs in FDA list with MMQ, co-occurrence counts and Levenshtein ratio
Early Prediction of Autism Spectrum Disorder Using Health Claims Data

Yu-Hsin Chen, MS1, Qiushi Chen, PhD1, Lan Kong, PhD2, Guodong Liu, PhD2
1The Pennsylvania State University, Department of Industrial and Manufacturing Engineering, University Park, PA, USA; 2The Pennsylvania State University, Department of Public Health Sciences, Hershey, PA, USA

Introduction

Autism spectrum disorder (ASD) is a lifelong neurodevelopmental disorder that affects every 1 in 54 children in the US.7 Certain medical conditions have found to be associated with ASD prior to the onset of behavioral symptoms, but no study has systematically examined the predictive value of the medical records in young children. Our objective is to explore the feasibility of predicting the ASD diagnosis using claims data in children between age of 18-30 months.

Methods

Using the MarketScan® Commercial Health Claims database from 2005-2016, we retrospectively created a study cohort of children with and without confirmed ASD diagnosis based on ICD-9 and ICD-10 codes (12,743 ASD and 25,833 non-ASD subjects identified in total). Independent predictor variables included demographic variables, number of emergency department visits, and numbers of healthcare encounters for each diagnosis and procedure category grouped by the Clinical Classifications Software. We compared Lasso logistic regression (LR) and random forest (RF) in predicting the ASD diagnosis given all available clinical information up to ages of 18, 24, and 30 months, respectively. We randomly sampled from the identified cohort to create a large balanced training dataset (N=20,000) with 10,000 ASD and non-ASD subjects, respectively, and an independent testing set (N=16,123) with a matched prevalence of ASD (1 in every 54) in the US. We repeated random sampling with 50 replications for model validation.

Results

At the age of 24-month, the LR and RF models showed the area under the receiver operating characteristic curve (AUROC) of 0.758 (95% Confidence Interval [CI]: 0.753-0.762) and 0.773 (0.770-0.777), respectively (Table 1). Compared with the LR model, RF model also showed higher area under precision-recall curve (AUPRC) and F1 score. The RF model achieved higher performance consistently in all other measures compared with the LR model. We also found that the prediction accuracy increased with the age of prediction, as more clinical information was available for model training at an older age. When the predictor variables were separated by outpatient and inpatient visits, the RF model for prediction at age of 24 months achieved a higher performance with an AUROC of 0.837, which translates into a promising improvement in sensitivity (40% vs. 38.8%), specificity (96.6% vs. 94.9%) and positive predictive value (PPV; 18.3% vs. 14.6%) compared with the existing screening tool.

Table 1. Prediction performance of Lasso logistic regression (LR) and random forest (RF) at 50% target sensitivity.

<table>
<thead>
<tr>
<th>Age at prediction</th>
<th>Model</th>
<th>AUROC</th>
<th>AUPRC</th>
<th>F1 score</th>
<th>Specificity, %</th>
<th>PPV, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base Case: 24-month</td>
<td>LR</td>
<td>0.758 (0.753, 0.762)</td>
<td>0.083 (0.079, 0.086)</td>
<td>0.174 (0.169, 0.179)</td>
<td>84.0 (83.5, 84.6)</td>
<td>5.7 (5.4, 5.9)</td>
</tr>
<tr>
<td></td>
<td>RF</td>
<td>0.773 (0.770, 0.777)</td>
<td>0.123 (0.117, 0.128)</td>
<td>0.223 (0.218, 0.228)</td>
<td>86.8 (86.4, 87.3)</td>
<td>6.7 (6.4, 6.9)</td>
</tr>
<tr>
<td>Younger: 18-month</td>
<td>LR</td>
<td>0.719 (0.715, 0.723)</td>
<td>0.054 (0.052, 0.056)</td>
<td>0.114 (0.110, 0.117)</td>
<td>78.6 (78.0, 79.2)</td>
<td>4.2 (4.1, 4.4)</td>
</tr>
<tr>
<td></td>
<td>RF</td>
<td>0.720 (0.716, 0.724)</td>
<td>0.057 (0.055, 0.059)</td>
<td>0.119 (0.115, 0.124)</td>
<td>79.1 (78.5, 79.8)</td>
<td>4.2 (4.1, 4.4)</td>
</tr>
<tr>
<td>Older: 30-month</td>
<td>LR</td>
<td>0.803 (0.800, 0.807)</td>
<td>0.132 (0.127, 0.137)</td>
<td>0.232 (0.223, 0.241)</td>
<td>91.0 (90.7, 91.4)</td>
<td>9.0 (8.2, 9.7)</td>
</tr>
<tr>
<td></td>
<td>RF</td>
<td>0.833 (0.830, 0.836)</td>
<td>0.209 (0.204, 0.215)</td>
<td>0.300 (0.295, 0.305)</td>
<td>95.4 (95.2, 95.7)</td>
<td>17.5 (16.7, 18.3)</td>
</tr>
</tbody>
</table>

Discussion

To the best of our knowledge, our study is the first to predict the ASD risk in young children based on longitudinal medical information from a large health claims database using machine learning models. Our findings have demonstrated the promise of leveraging real-world health claims data to monitor the risk of ASD among very young children and to identify those at high risk for targeted screening, diagnosis and early interventions.

References

Data Coordination for Multi-Site Clinical Trials Using the REDCap Application Programming Interface

Alex C. Cheng, Ph.D., MEM, Mark McEver, Francesco Delacqua, Adam Lewis, MS, Patrick Newman, MBA, Paul A. Harris, Ph.D.
Vanderbilt University Medical Center, Nashville, TN

Introduction

Coordinating data collection, transfer, and harmonization is a challenging task for multi-site clinical trials. Data coordinating centers (DCC) devote much effort to making sure data is collected uniformly across sites. Additionally, study sites are often wary of entering data directly into an electronic data capture system at a DCC outside their hospital firewall without first having the opportunity to validate and de-identify the data. Due to these challenges, there has been increased interest among groups such as the NIH Collaboratory and the Trial Innovation network for a decentralized model for multi-site trial data collection. We have designed a framework that gives sites control of their own data collection while addressing some inefficiencies of multi-site clinical trial data coordination using REDCap as an engine for data interoperability. REDCap (Research Electronic Data Capture) is a secure, web-based software platform designed to support data capture for research studies in use at nearly 5000 institutions in 141 countries. Its ubiquity among medical centers conducting clinical research use makes it an ideal platform for coordinating multi-site data collection.

Methods

Figure 1 illustrates the data flow for a multi-site trial that collects EHR data and uses this framework. At each site, REDCap uses the HL7 Fast Healthcare Interoperability Resources (FHIR) application programming interface (API) to extract data from the EHR into REDCap. Sites’ study personnel use REDCap’s Clinical Data Interoperability Services (CDIS) module to map data elements available through their EHR FHIR APIs to data elements in the study data dictionary. An external module called API Sync allows for automatic data pushes between REDCap projects with compatible data fields. An authenticated user generates an API key for the DCC REDCap project, which sites use to push data to the DCC. This allows data to flow automatically and securely from the sites to a DCC at a specified frequency.

Results

We are applying this framework to two ongoing multi-site studies where Vanderbilt University Medical Center (VUMC) is the DCC. These studies have approved waivers of informed consent. So far, for one of these studies, two sites have transferred 883,133 data elements for 972 study participants to VUMC using API Sync. Statisticians are able to analyze, in real-time as data are entered by the sites, every participant record from the sites, with EHR data collected through the FHIR API. All data is de-identified and date shifted by REDCap before being sent to the DCC. Currently, 99 institutions have downloaded API Sync from the REDCap External Module Repository.

Discussion

This data collection and transfer framework offers several advantages over traditional methods for data coordination in multi-site clinical trials. The data dictionary distributed to the sites allows the DCC to provide structure for study data while giving site control over their data collection. Using API Sync to transfer data streamlines the process of having to download, de-identify, reformat, send, merge, and harmonize data at the DCC. These processes can all be done within REDCap and scheduled to occur on a regular frequency. The use of REDCap APIs is enabling an automated data pipeline that reduces study personnel burden and increases the accuracy and efficiency of data collection and coordination.

Funding: This work is supported by NIH/NLM contract #75N97019P00279
An eConsent Process Framework to Optimize and Standardize Development, Deployment, and Management of Electronic Informed Consents

Benjamin Chisum, BS¹, Kevin M. Johnson, MBA¹, James Ruesch, MBA, CCRP¹, Patti Spencer, BS¹, Pallavi Ranade-Kharkar, MS, PhD, FAMIA¹
¹Intermountain Healthcare, Salt Lake City, UT

Introduction
Research suggests that the delivery of a research informed consent via electronic means (eConsent) can improve participant’s informed consent experience, ease of access, broaden opportunity to participate in trials, improve enrollment potential, and substantially lower the cost of human capital. However, little guidance is available regarding the development of a successful standardized eConsent delivery system, especially across large research systems where numerous clinical trials are conducted. Guidance regarding deployment and management of each study’s unique eConsent process is equally sparse. The process of developing an effective eConsent tailored to the unique requirements of each study and subsequent deployment is complex and many unforeseen technical, operational, and user-related issues may arise. The objective of this project is to propose a framework to optimize, standardize, and disseminate guidelines and best practices for development, deployment, and management of eConsents.

Method
Intermountain Healthcare has developed and deployed more than 100 eConsents since 2015, utilizing multiple platforms. The primary platform used was REDCap. We applied this experience and lessons learned to develop a generalizable eConsent framework. The framework provides instructions and guidelines for the development, deployment, and management life cycle of an eConsent. Development of the framework occurred over the course of many instances of eConsent development for many research departments. Each department’s input helped further grow the framework. As each eConsent was deployed, the framework was subsequently enhanced, optimized, and standardized. The framework spans the process end-to-end from eConsent platform requirements all the way to archival methodology. This framework is a living application that allows active updates to grow the comprehensive eConsent guidelines, experience, and standardization.

Results and Discussion
The eConsent framework is comprised of multiple sections, including: ‘System/software specification,’ ‘Compliance,’ ‘Development,’ ‘Deployment,’ ‘Management,’ and ‘Study closeout/archival.’ Each section includes sub-sections that further outline sub-processes. For example, sub-sections under ‘Development’ include ‘Deployment method,’ ‘Design,’ ‘Communication modality and language,’ ‘Method for subjects to ask questions,’ ‘Legal certification and verification,’ and ‘Return completed consent to subject.’ We recently applied the framework to two high-enrolling studies. One of the studies with a participant enrollment goal of 9,000 projected that the study would require 20 coordinators to facilitate traditional informed consenting. Strategies from the eConsent framework lead to successful deployment of an eConsent that allowed the study to operate with only three study coordinators. The study is subsequently saving more than $1,000,000 in annual salary and benefit expenses. Another study is on-track to enroll 500,000 subjects over the course of five years. This goal would be unfeasible utilizing traditional coordinator-based consenting. Further, a vended eConsent product for this enrollment goal would cost at least $10,000,000 (at $20 per patient per eConsent). During COVID, the guidelines of the eConsent framework allowed for successful and rapid eConsent development and deployment for extremely fast-moving trials.

Conclusion
Development of an optimized and standardized eConsent framework can enable successful and efficient deployment and management of eConsents. Detailed, actionable, and sharable knowledge regarding the lifecycle of an eConsent can improve growth, efficiency, and success of future eConsents. We would like to further develop a platform that is accessible to the global research community and allows for collaboration towards growth of the body of knowledge.

References
Usability of Clinical Decision Support System for Nursing Care Planning in Palliative Care: Heuristic Evaluation

Hwayoung Cho, PhD, RN1, Gail Keenan, PhD, RN, FAAN1, Olatunde Madandola, MPH, RN1, Fabiana Dos Santos, MSN1, Tamara Macieira, PhD, RN1, Ragnhildur Bjarnadottir, PhD, MPH, RN1, Karen Priola, MS1, Karen Dunn Lopez, PhD, RN2

1College of Nursing, University of Florida, Gainesville, FL; 2College of Nursing, University of Iowa, Iowa City, IA

Introduction

A clinical decision support system (CDSS) is a persuasive tool aimed at supporting evidence-based practice by providing the right information in the right format at the right time to improve patient outcomes at the point of care. In response to the lack of CDSS tools for nurses in palliative care, our study team developed an electronic nursing Care Planning System (CPS) that facilitates evidence-based decision making for nurses caring for end of life patients.1 Our CPS is an electronic nursing documentation and care planning prototype system that can be interfaced with electronic health records (EHRs). Despite the proliferation of technology (EHRs/CDSS) in today’s health care, many systems have not been rigorously evaluated. We aimed to evaluate the usability of the CPS from expert perspectives, before we test the efficacy of the CPS in a randomized controlled trial (RCT) with a nationally representative sample of 220 registered nurse (RN) subjects.

Methods

We conducted a heuristic evaluation with experts in informatics to identify violations of usability principles. Using purposive sampling, six experts were invited using email. Qualifications of the experts included having a doctorate degree in informatics and training in human-computer interaction. Each expert was provided an orientation (PowerPoint video) about how the CPS worked and two use cases that would require users to adjust the plans to the unfolding context. Each expert was then asked to independently interact with the CPS and to complete a heuristic evaluation checklist based on Nielsen’s 10 heuristics.2 The overall severity for each heuristic item was rated on a scale of 0 (no problem) to 4 (usability catastrophe), and space was provided for experts to add comments. Because our upcoming clinical trial will test evidence-based suggestions using three types of formats (text, table, graph), Aesthetic and Minimalist Design heuristic was evaluated for each format. All experts’ comments about usability problems on the heuristic evaluation checklist were synthesized by two health information technology usability experts (HC&KDL). Discrepancies in coding data according to the usability factors of Nielsen’s 10 heuristics were discussed about usability problems on the heuristic evaluation checklist. The mean scores of the overall severity of the identified heuristic violations ranged from 0.66 (Flexibility and Efficiency of Use) and 2.00 (User Control and Freedom/Error Prevention), in which scores closest to 0 indicate a more usable system. The mean scores and sample comments per heuristic are listed in Table 1. To maintain User Control and Freedom, experts suggested the ‘undo’ function not be limited and to give users the ability to fix both current/expected nursing outcome scores. To improve Recognition Rather than Recall, experts recommended clicking the ‘undo’ button to see what was undone should be recognizable within the CPS. In response to the usability factor Match between System and the Real World, experts pointed to the opposite direction in pain scale scores used in our system when compared with those commonly used in clinical practice.

Results

Mean scores of the overall severity of the identified heuristic violations ranged from 0.66 (Flexibility and Efficiency of Use) and 2.00 (User Control and Freedom/Error Prevention), in which scores closest to 0 indicate a more usable system. The mean scores and sample comments per heuristic are listed in Table 1. To maintain User Control and Freedom, experts suggested the ‘undo’ function not be limited and to give users the ability to fix both current/expected nursing outcome scores. To improve Recognition Rather than Recall, experts recommended clicking the ‘undo’ button to see what was undone should be recognizable within the CPS. In response to the usability factor Match between System and the Real World, experts pointed to the opposite direction in pain scale scores used in our system when compared with those commonly used in clinical practice.

Conclusion

The results of this heuristic evaluation will be used to guide improvements in our CPS design/interface including content, organization and workflow. These refinements will be made in our current RCT with 220 RNs in which we are testing the optimal CPS format and decision time, and evaluating simulated patient outcomes related to the CPS formats. Rigorous testing of the CPS using heuristic evaluations can enhance the likelihood of adoption of the CPS to diverse nurses in real-world practice.

Acknowledgements This study is supported by the National Institutes of Health, National Institute of Nursing Research (R01 NR018416-01).

References


Table 1. Mean Scores and Sample Comments from Heuristic Evaluation

<table>
<thead>
<tr>
<th>Nielsen’s 10 usability heuristics</th>
<th>Mean (SD)</th>
<th>Sample Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visibility of System Status</td>
<td>1.80 (1.21)</td>
<td>Unclear if care plan icons are clickable</td>
</tr>
<tr>
<td>Match between System and the Real World</td>
<td>1.60 (1.09)</td>
<td>Opposite direction of scoring in Pain scale</td>
</tr>
<tr>
<td>User Control and Freedom</td>
<td>2.00 (1.09)</td>
<td>Limited ‘Undo’ functionality</td>
</tr>
<tr>
<td>Consistency and Standards</td>
<td>1.16 (1.16)</td>
<td>Unclear of formatting standards referred</td>
</tr>
<tr>
<td>Help users Recognize, Diagnose, and Recover from Errors</td>
<td>1.66 (1.63)</td>
<td>Error message is not informative</td>
</tr>
<tr>
<td>Error Prevention</td>
<td>2.00 (1.09)</td>
<td>Need warning message when clicking ( )</td>
</tr>
<tr>
<td>Recognition Rather than Recall</td>
<td>1.83 (1.16)</td>
<td>Need HELP function on how CPS works</td>
</tr>
<tr>
<td>Flexibility and Efficiency of Use</td>
<td>0.66 (1.03)</td>
<td>Suggested hyperlinks, alphabetical index</td>
</tr>
<tr>
<td>Help and Documentation</td>
<td>1.83 (0.98)</td>
<td>Not visually appealing from similar blues/grey shades</td>
</tr>
</tbody>
</table>

The CPS was synthesized by two health information technology usability experts (HC&KDL).
A Quantitative Exploration of Nurses’ Numeracy and Understanding of Predictive Models

Insook Cho, PhD1*, Mi Ra Song, MS2, Shim So Yun, MS2, Dan So Young, MS2
1Nursing Department, Inha University, Incheon, Republic of Korea; 2Department of Nursing, Samsung Medical Center, Seoul, Republic of Korea

Introduction

The rapid expansion of electronic health-care predictive models stemming from the increasing availability of electronic health-record data is introducing new challenges and opportunities that are not covered by earlier health-care models, algorithms, and tools. One concern is that clinicians may not receive the requisite professional training for utilizing the emerging predictive modeling paradigm in health care. The concerns range from clinicians’ and managers’ poor understanding of concepts in probability, statistics, and heuristics to the health-care professional curricular and training. Given the novelty of predictive models to nurses, this study aimed to quantitatively explore nurses’ understanding and their perceptions about probability-based information of predictive models.

Methods

Sixty nurses working at six medical-surgical nursing units in a tertiary academic hospital located in Seoul were recruited in 2020. We applied a stratification sampling framework to the participating nurses according to their level of experience at the unit: novice (<2 years), junior (2–5 years), and senior (>5 years). The survey instrument included educational modules about predictive models, sensitivity, specificity, and confidence intervals (CIs) [1]. Follow-up questions tested the abilities of the participants to interpret these characteristics with both verbatim and gist knowledge. Each module presented the statistical concept in the context of a weather prediction, which was chosen to remove any potential cognitive or affective influences common in medical decision making. Verbatim knowledge is the ability to interpret specific number associated with the model performance characteristics described in the question stems. Gist knowledge is the ability to discriminate the central meaning qualitatively. Objective numeracy and subjective numeracy were also assessed using the short form of the Numeracy Understanding in Medicine instrument (S-NUMi), the Subjective Numeracy Scale (SNS), and an objective numeracy scale [2].

Results

Among 60 nurses, the mean correct responses were 95.0% for sensitivity, 93.9% for specificity, and 68.9% for CIs. Verbatim interpretation was significantly higher than gist knowledge for sensitivity ($\chi^2=8.57, p=0.0061$), but not for specificity and CIs ($\chi^2=4.90, p=0.0535; \chi^2=0.54, p=0.4642$). Scores on each discrete choice experiment tasks were high in each category (Table 1). The overall knowledge score was 85.9%, which was not associated with the scores of subjective and objective numeracy measures. The scores of S-NUMi, SNS, and objective numeracy scale were 7.00±0.88 (mean±SD), 4.65±0.83, and 6.17±0.92, respectively.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall knowledge</td>
<td>85.9</td>
</tr>
<tr>
<td>Sensitivity</td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>95.0</td>
</tr>
<tr>
<td>Verbatim</td>
<td>91.8–98.2</td>
</tr>
<tr>
<td>Gist</td>
<td>86.7</td>
</tr>
<tr>
<td>DCE task</td>
<td>78.1–95.3</td>
</tr>
<tr>
<td>Specificity</td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>93.9</td>
</tr>
<tr>
<td>Verbatim</td>
<td>90.4–97.4</td>
</tr>
<tr>
<td>Gist</td>
<td>96.7</td>
</tr>
<tr>
<td>DCE task</td>
<td>92.1–100</td>
</tr>
<tr>
<td>95% CIs</td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>68.9</td>
</tr>
<tr>
<td>Verbatim</td>
<td>62.1–75.7</td>
</tr>
<tr>
<td>Gist</td>
<td>50.0</td>
</tr>
<tr>
<td>DCE task</td>
<td>37.4–62.7</td>
</tr>
</tbody>
</table>

Scores are reported as mean percentages with 95%-confidence intervals. DCE means discrete choice experiment.

Conclusion

These results suggest that nurses can interpret quantitative performance measures of predictive models with high accuracy but have considerably lower when interpreting CIs. Their numeracy scores were moderate-to-high level but was not related to ability to interpret performance measures.

References

*This study was supported by grants from the Ministry of Trade Industry and Energy (20004861) and the NRF of Korea (2019R1A2C2007583).
Designing a Data Quality Characterization Tool for Fitness Tracker Data

Sylvia Cho, MHS, MA¹, Karthik Natarajan, PhD¹
¹Department of Biomedical Informatics, Columbia University, New York, NY

Abstract
In this study, design requirements of a data quality (DQ) characterization tool for fitness tracker data was identified through findings from previous studies on DQ challenges and dimensions, literature review to find fitness-for-use measures, and interviews with stakeholders. Two design requirements were identified: (1) the tool should be able to characterize data in terms of DQ dimensions and corresponding measures, and (2) contextual information such as metadata and demographic data should be provided.

Introduction
Lack of trust in the quality of data is a barrier to using electronic health data. As DQA is a time-consuming and difficult task, a tool that supports DQA would facilitate the use of data. DQ is commonly defined as fitness-for-use which means that the quality of data is dependent on users’ needs or a specific task. While existing tools provide descriptive statistics or exploratory data analysis that are independent of any specific use case (e.g., percent missingness in a column), it rarely focuses on fitness-for-use measures that are specific to a research question. The aim of this study was to identify design requirements of a DQ characterization tool for fitness tracker data focusing on fitness-for-use measures specifically for data completeness.

Methods
Design requirements were identified through three methods/sources: (1) findings of previous studies on DQ dimensions and DQ challenges [1,2], (2) literature review to identify fitness-for-use data completeness measures that are commonly used in research studies utilizing fitness tracker data, and (3) semi-structured interviews on the current practices of data completeness assessment and challenges of determining the fitness-for-use of a dataset. Thematic analysis was conducted on the transcript.

Results
Two major informational needs emerged from the methods and sources we used. The results are presented in Table 1.

Table 1. Informational Needs and Design requirements Identified through previous studies, literature review, and interview

<table>
<thead>
<tr>
<th>Semi-structured Interview (Interview with 3 target end users)</th>
<th>Findings from Previous Studies (DQ Dimensions, Challenges)</th>
<th>Literature Review (Fitness-for-use Measures)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. <strong>Best practices and standards are important</strong></td>
<td>Breadth completeness[2]</td>
<td>NA</td>
</tr>
<tr>
<td>Interviewees confirmed that the lack of best practices on DQA can be a challenge to researchers unfamiliar with emerging data types. The interviewees also confirmed that the fitness-for-use measures found from literature makes sense.</td>
<td>Density completeness[2]</td>
<td>(Example)</td>
</tr>
<tr>
<td></td>
<td>Does the dataset have all variables I need to answer the research question?</td>
<td>Number of valid days per certain period</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Valid day criteria</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Has step count above a certain threshold</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Has hours of data above a certain threshold</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Hour with data can be calculated based on:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>o Threshold to determine wear time vs. non-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>wear time</td>
</tr>
<tr>
<td>2. <strong>Importance of Contextual information</strong></td>
<td>Interview findings align with the importance of metadata (e.g., device type, data collection period) due to data heterogeneity problem[1]</td>
<td>NA</td>
</tr>
<tr>
<td>• Device type</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Time of data collection</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Demographics</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Based on these findings and the authors’ knowledge on DQA, the prototype of the tool included the following features:
1. **Users can define density completeness that fits their use case**: Based on a previous study, the tool should be able to characterize DQ in terms of breadth and density completeness and its corresponding measures.[2]
2. **A data summary on the cohort that meets the completeness definition**: The tool calculates the total number of subjects meeting the completeness criteria for each data types (e.g., step count, heart rate) and summarizes metadata such as device type and data collection period, and also descriptive statistics on demographics.

Conclusion
In this preliminary study, design requirements of a DQ characterization tool with a goal to support fitness-for-use was investigated and applied to the design of the tool. A future study will be conducted to evaluate whether a DQ characterization tool focusing on fitness-for-use measures is more useful to researchers than an intrinsic DQ characterization tool.

References
Dynamic Phenotyping of Multi-Modal COVID-19 Endotypes

Sutanay Choudhury\textsuperscript{1}, Khushbu Agarwal\textsuperscript{1}, Colby Ham\textsuperscript{1}, Pritam Mukherjee\textsuperscript{2}, Siyi Tang\textsuperscript{2}, Sindhu Tipirneni\textsuperscript{3}, Chandan Reddy\textsuperscript{3}, Suzanne Tamang\textsuperscript{2}, Robert Rallo\textsuperscript{1}, Veysel Kocaman\textsuperscript{4}
\textsuperscript{1}Pacific Northwest National Laboratory; \textsuperscript{2}Stanford University; \textsuperscript{3}Virginia Tech.; \textsuperscript{4}John Snow Labs

Introduction

We aimed to dynamically characterize different multi-modal endotypes (i.e., subtypes) of COVID-19 patients and the effectiveness of hospital-based treatments, over the course of the pandemic. Our analysis found that among COVID-19 patients that experienced a poor outcome, the average LOS reduced after June, while remaining consistent for admissions that did not experience a poor outcome. Also, that clinicians were able to more effectively treat severe COVID-19 cases – or at least specific endotypes of complex, high utilization cases – as the outbreak progressed.

Methods

We conduct our study using a deidentified version of the STARR-OMOP dataset containing both structured electronic healthcare records and free form clinical notes from Stanford Medicine. We represent each patient stay in the hospital as a sequence of sets over time (subsequently referred to as patient trajectory), in which each set represents a snapshot of the patient state comprising diagnosis codes, drug orders and laboratory measurements aggregated over time. Each admission was assessed to determine if a poor COVID-19 related outcome occurred, which was true in the event of ventilator support, ICU admission or inpatient death. To characterize baseline risk factors, we processed each patient’s clinical notes in the first 24 hours of their visit. We processed the clinical notes using the entity resolver model from SparkNLP framework to translate the clinical notes to another temporal set-based data representation. We used a frequent pattern mining method\textsuperscript{1} to discover frequent endotypes among COVID-19 cases.

Results

We show descriptive statistics for our COVID-19 cohort in Figure 1(a). Figure 1(b) shows the monthly ranking of baseline risk factors extracted with NLP. Figure 1(c) compares the distribution of LOS over time for patients with a poor COVID-19 related outcome to the rest of the cohort.

<table>
<thead>
<tr>
<th>Number of hospitalizations</th>
<th>454</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean Length of stay (SD)</td>
<td>8.43 / 10.04</td>
</tr>
<tr>
<td>Mean Age (SD)</td>
<td>48.8 (21.3)</td>
</tr>
<tr>
<td>Race Counts [African-American, Asian, Hispanic or Latino, White, Native Pacific Islander, Other]</td>
<td>(21, 45, 221, 100, 19, 48)</td>
</tr>
<tr>
<td>Outcomes [Ventilation, ICU, Mortality]</td>
<td>(50/404, 43/411, 24/430)</td>
</tr>
</tbody>
</table>

Figure 1: (a) Summary statistics of the dataset (b) patient risk factors at time of admission over the course of epidemic (c) patient LOS for severe vs non-severe patients over time.

Acknowledgements

This research was supported by Pacific Northwest National Laboratory LDRD program and the Stanford Center for Population Health Sciences.

References

Title: Using Deep Neural Network Binary Classifiers to Infer MeSH Publication Types

Victor Cid, MS and James Mork, MSc, National Library of Medicine, Applied Clinical Informatics Branch, National Institutes of Health, Bethesda, MD

Introduction: We use Deep Learning (DL) models to label PubMed’s MEDLINE citations using a 59-Publication Type (PT) subset of MeSH Publication Types\(^1\) used by NLM indexers. Our training corpus is limited to metadata in MEDLINE citation records. We trained binary classifiers using modern Transformer-based deep learning models with labeled data from PubMed’s 30 million citations, and combine these classifiers on an application that can process tens of thousands of citations daily. We developed a method to carefully engineer training examples in order to maximize the performance of our DL models.

Methods: We use MEDLINE citations to train BERT and DistilBERT\(^2\) models using transfer learning for binary text classification. Previous efforts have used multi-class classifiers\(^3\). There is great variability in the data available for different PTs, ranging from a few hundred to over a million records. We address data balance in a per-class fashion. The corpus data contains over 20 features, but through experimentation we selected Title, Abstract, Authors and Journal Unique ID as best class predictors. We developed a method to prepare training datasets to maximize model performance. The 59 trained models were combined to produce multiple labels for each document, excluding prohibited co-occurrences, and results with scores over 0.93 (determined experimentally), were included in the final output PT labels.

Results: This work is ongoing. We achieved high recall and F1-scores (both 0.84 to 0.99) for all individual trained models, but the application combining all 59 models achieved only a 61% labeling accuracy. Pruning lower-performing models increased the application performance up to about 80% accuracy. Our data engineering approach proved to enhance the application’s inference accuracy by about 15% compared to random sampling, and reduce false positives by about the same amount. DistilBERT proved to be about 50% faster to train and about 90% faster to perform class inferencing than BERT, while conveying very similar training scores. The inference time is particularly important as the application must process up to about 20,000 citations daily, which even with DistilBERT can take hours to complete in our experimental setup. In an effort to achieve better application accuracy, we are both, enhancing the training of classifiers via varying training parameters, and pruning models from the application that we cannot further improve with the available data when necessary. Furthermore, this work is also providing insight on the value of different MEDLINE data features to predict publication types.

Acknowledgements: This work was supported by the Intramural Research Program of the U.S. National Institutes of Health (NIH), National Library of Medicine (NLM), and Lister Hill National Center for Biomedical Communications (LHNCBC).

References

National COVID Cohort Collaborative (N3C) Case-control Buddies

G. Marshall Clark¹, Adam M. Lee, MBA¹, Emily R. Pfaff, PhD¹, Kristin Kostka, MPH² ³, Matvey B. Palchuk, MD⁴, Lora Lingrey, BS⁵, Michele Morris, BA⁵, Robert Miller, MS² ⁶¹University of North Carolina at Chapel Hill, Chapel Hill, NC, ²Real-World Solutions, IQVIA, Cambridge, MA ³The OHDSI Center at The Roux Institute, Northeastern University, Portland, ME ⁴TriNetX, Cambridge, MA, ⁵University of Pittsburgh, Pittsburgh, PA, ⁶Tufts Clinical and Translational Science Institute, Boston, MA

Introduction

The National COVID Cohort Collaborative (N3C) required control matches for COVID-19 cases identified by the N3C Phenotype team. The N3C Phenotype team created a case-control matching system that algorithmically assigns persistent case-controls coined as phenotype buddies for the data submissions to the N3C enclave.

Methods

The N3C phenotype 3.0 introduces specific SQL in multiple code variants that create data structures and match controls to the phenotype cases. The code creates tables locally at each site that hold all cases and controls, specifically a permanent data structure, N3C-CONTROL-MAP, table that houses the cases and control population and variables: age group, sex, race, and ethnicity for cases and their buddies. Ages are grouped in 5-year bins, samples are (0-4), (5-9), (85-89), (90+). This table allows the phenotype to maintain persistent buddies per case, while the phenotyping algorithm dynamically checks both cases and controls for cohort eligibility. Cases that are no longer eligible become potential controls, and controls classified as new cases are removed as buddies, and new buddies are assigned. The phenotype progressively matches controls based on variable sets, with each group achieving a matching level as indicated in Table 1, with level 4 being the best match. The data structure and matched buddies are bundled in each site’s data payload to the N3C enclave providing a back-up to the N3C-CONTROL-MAP in case of data loss or corruption. The phenotype SQL code is publicly available at: https://git.io/JqZHi

Results

The N3C Phenotype team’s case-control buddy system matches 90.87% of cases with viable controls on age group, sex, race, and ethnicity. Exact age is the most unmatched factor of the individual variables, with only 34.62% of all cases matching, but creating the age groups increased that match percentage to 98.22%. The remaining variables checked independently are as follows: sex (99.99%), race (97.00%), and ethnicity (96.96%).

<table>
<thead>
<tr>
<th>Match Level</th>
<th>Buddy 1</th>
<th>Buddy 2</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>4 (AGE, SEX, RACE, ETHNICITY)</td>
<td>45.42%</td>
<td>45.44%</td>
<td>90.87%</td>
</tr>
<tr>
<td>3 (AGE, SEX, RACE)</td>
<td>0.72%</td>
<td>0.69%</td>
<td>1.42%</td>
</tr>
<tr>
<td>2 (AGE, SEX)</td>
<td>1.78%</td>
<td>1.83%</td>
<td>3.61%</td>
</tr>
<tr>
<td>1 (SEX)</td>
<td>2.07%</td>
<td>2.04%</td>
<td>4.10%</td>
</tr>
<tr>
<td>No Match</td>
<td>0.01%</td>
<td>0.00%</td>
<td>0.01%</td>
</tr>
</tbody>
</table>

Table 1 Control Buddy percentage by match level

Discussion

The current phenotype version 3.1 supports two buddies per case; revisions are underway to scale the number of buddies to any number of case-control matches as requested by N3C and deliverable by each site. While the current iteration of the N3C Phenotype buddy algorithm matches four variables, the code is extensible to add additional matching variables, such as secondary race, geographical region, co-morbidity, or other suspected confounders. Albeit, the extensibility is not ad hoc or site-specific and is determined by N3C. The N3C phenotype buddy system only provides suitably matched controls of non-exposure patients defined by the N3C phenotype; it does not provide study-specific analysis, odd ratios, or insights into efficiency or collider bias¹.

Conclusion

Overall, the N3C Phenotype team developed and implemented a case-control solution for the advancement of COVID-19 research. This phenotype and the buddy matching algorithm allow N3C and collaborators to analyze controls on the over 900,000 COVID19 positive cases currently in the N3C enclave.

References

Healthcare System Performance on Unsafe Medication Orders in the 2019 CPOE Evaluation Tool

Zoe Co, BS1, Lisa P. Newmark, BA2, David C. Classen, MD, MSc3, Diane L. Seger, RPh2, Melissa Danforth, BA4, Jessica M. Cole, BS3, David W. Bates, MD, MSc1,2,5

1Department of General Internal Medicine, Brigham and Women’s Hospital, Boston, MA, 2Clinical and Quality Analysis, Mass General Brigham, Somerville, MA, 3The University of Utah, Salt Lake City, UT, 4The Leapfrog Group, Washington D.C., 5Harvard Medical School, Boston, MA

Introduction: Since the HITECH Act was passed in 2009, electronic health record (EHR) implementation has grown widespread.1 Hospitals in the United States use The Leapfrog Group’s CPOE Evaluation Tool to assess the ability of their EHR to alert on common and serious prescriber errors. The tool provides hospitals with an overall percentage score of unsafe orders detected, as well as individual order category scores. Past analyses of the tool have reported that EHR implementation varies greatly within vendors and even between hospitals within the same healthcare system.2 In this study, we report on the variability in performance of hospitals within four healthcare systems in terms of their overall percentage score and individual order checking category scores.

Methods: A dataset containing the 2019 CPOE Evaluation Tool results was extracted for analysis and was linked to the American Hospital Association Annual Survey from 2009 to 2016 to identify hospitals that were part of a healthcare system. Next, we selected four healthcare systems that contained 20 or more hospitals from the dataset and identified them as Systems A-D. Within each of these healthcare systems, all the hospitals used the same combination of EHR vendor and medication reference database. For each healthcare system, we reported on the range of order category scores their hospitals received within in the test, as well as their overall performance.

Results: The complete dataset contained 119 hospitals across four healthcare systems, with two different combinations of EHR vendor and medication databases (Table 1).

There was variability in hospital performance within each healthcare system, with System C having the largest range of overall scores (Table 1). In each system, there were order categories that showed great variability in performance. For example, in System A, the drug laboratory, drug age, and drug diagnosis order category scores ranged from 0% to 100%, suggesting that some hospitals in this healthcare system had these alerts completely turned off. A similar case also occurred in System D, where scores ranged from 0% to 100% in the therapeutic duplication, drug age, and drug monitoring order categories.

The analysis also identified differences in how hospitals within the same healthcare system respond to the same medication test orders they received in the test. For example, 16 hospitals in System A responded differently to the same drug monitoring order that involved tobramycin. Of the 16 hospitals, 8 received an alert to monitor tobramycin levels while the 8 other hospitals did not receive the alert.

Conclusion

The results of these analyses show the variability of EHR implementation at the facility level. By having some areas of clinical decision support completely turned off, there is potential for adverse drug events to occur. This study reveals important areas of improvement and reinforces the importance for healthcare systems to repeatedly assess the performance of their hospitals’ EHRs.

References


Table 1: Overall performance of the four healthcare systems.
Comparative Analysis of Telemedicine Services Before and During Pandemic

Wanting Cui, MA, Joseph Finkelstein, MD, PhD
Icahn School of Medicine at Mount Sinai, New York, NY

Introduction

Since the start of the COVID-19 pandemic, the number of telemedicine services increased significantly due to the ‘stay at home’ order in New York City. The goal of this study is to identify demographic groups of patients who received telemedicine services, and to analyze how different demographic groups’ telehealth usage patterns change throughout the course of the pandemic.

Method

The dataset was generated by querying electronic health records at the Mount Sinai Health System to identify all patients who used telemedicine services between January 2019 and December 2020. Patients’ demographics, diagnoses and medical history are the variables in the dataset. We calculated patients’ Charlson comorbidity index based on patients’ medical history (ICD10 codes) and age. We divided the dataset into 2 sections: prior to the pandemic (2019/1/2 – 2020/2/28) and during the pandemic (2020/3/1 – 2020/12/31). All analyses were performed in Python (Python version 3.7). We performed T-Tests for numerical variables and Pearson Chi-square tests for categorical variables. All statistical tests were two-sided, with p<0.05 being considered statistically significant.

Result

There were 263,051 telemedicine sessions and 169,178 unique patients in the dataset. There was a drastic increase in telemedicine service starting 03/2020 and reached the peak in 05/2020 (Figure 1). The number of daily telemedicine sessions reached a stable but high level in 07/2020.

Figure 1. Average weekday telemedicine cases by month.

Prior to the pandemic, there were significantly more young patients and female patients using telemedicine services. However, these differences only reflected in adult patients, but not in children. The average age for adults was 42.0 years old and 7.9 years old for children prior to the pandemic. The average age increased to 49.9 years old for adults and remained constant at 7.7 years old for children during the pandemic. In addition, prior to the pandemic, only 33.1% of adult patients were male. The number increased to 39.3% in 2020. Although, the proportion of adult African American telemedicine users increased from 9.2% to 14.1%, there were significantly (p <0.001) more White telemedicine users (58.1%) than African American users. Furthermore, the average adult patients’ comorbidity scores also increased significantly (prior: 0.7, post: 1.4) during the pandemic.

Conclusion

The COVID-19 pandemic changed the landscape of telemedicine drastically. The average age of patients increased since the pandemic and there were significantly more White patients using the service than African American patients. In future studies, we plan to study the accessibility of telemedicine to older adults and the disparities of telemedicine usage between different races. Thus, future analyses of telemedicine are warranted.
Machine Learning for Time to Event Biomedical Research

Ioana Dančiu1,2,*, Greeshma Agasthya1, Janet Tate3,4, Amy Justice3,4,5, Ben McMahon6, Olga Ovchinikova1

1 Oak Ridge National Laboratory, Oak Ridge, TN, USA; 2 Department of Biomedical Informatics, Vanderbilt University, Nashville, TN, USA; 3 Veterans Affairs Connecticut Healthcare System, West Haven, CT, USA; 4 Yale School of Medicine, New Haven, CT, USA; 5 Yale School of Public Health, New Haven, CT, USA; 6 Theoretical Biology Group, Los Alamos National Laboratory, Los Alamos, NM, USA; *Corresponding author

Introduction

In recent years, access to an increasing volume and diversity of biomedical data, as well as advances in computational capabilities have generated significant interest in application of machine learning (ML) for biomedical applications (1). In this abstract, we present a domain aware predictive modeling approach for prostate cancer disease progression using machine learning that considers time-to-event with censoring.

Methods

Data: Our cohort consisted of patients with early-stage prostate cancer at diagnosis from the Department of Veterans Affairs (VA) Corporate Data Warehouse (CDW). This study has been approved by the VA and Oak Ridge National Laboratory (ORNL) Institutional Review Boards (IRB). 64 predictors were measured at or before primary prostate cancer diagnosis, and fell in two categories: (i) A set of independent variables abstracted from the CDW cancer registry characterizing patient demographics (age at diagnosis, race) and disease staging (Gleason score clinical, stage group ajcc and SEER summary stage); (ii) Longitudinal prostate specific antigen (PSA) values, aggregated into minimum, maximum, average, density, standard deviation values across 6 months intervals and also across a 5-year period prior to diagnosis. The outcome was a composite consisting of cancer related death from the National Death Index and registry documentation, PSA> 50, and inpatient and outpatient ICD codes indicative of metastatic disease. We censored all patients one year after their last PSA value.

Computational methods: We used xgboost with accelerated failure time (2) that maximized the log likelihood of a decision tree ensemble to calculated time to event survival in months. The model training used 80% of the dataset set aside at the beginning. The remaining 20% was used solely for independent testing of the trained model to compute the performance metric.

Metric: To determine the performance of our algorithms, we used a slightly modified version of Harrell’s c-statistic (3) that uses the time to event predictions instead of the hazard to generate the list of concordant pairs and ties.

Results

Our dataset consisted of 135,974 patients with early-stage prostate cancer diagnosed between 2002 and 2016, 108,260 in the training dataset and 27,714 in the validation set. 11,781 had the composite outcome before censoring in the training dataset, with 3,011 in the test. The median survival for the training dataset was 94 months 95% CI [94, 94] and 94 months 95% CI [93, 95] in the test. The modified Harrell’s c-statistic on the test dataset was 0.63.

Conclusions

Machine and deep learning methodologies that can capture the full complexities of datasets are both needed and underrepresented in the literature and current biomedical practice. Our research demonstrates that adapting existing machine learning technologies to account for time to event outcomes with censoring is feasible.

Acknowledgements

This work was supported by Department of Veterans Affairs, Office of Research and Development, Million Veteran Program MVP000 and MVP017. This manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy.

References

Disparities in Coded and Imputed Post-Traumatic Stress Disorder and Self-Harm Among US Veterans

Sharon E. Davis, PhD; Praveen Kumar, MS; Nicolas R. Laune, MS; Sharidan K. Parr, MD; Daniel Park, MS; Michael E. Matheny, MD, PhD; Gerardo Villarreal, MD; George Uhl, MD; Yiliang Zhu, PhD; Mauricio Tohen, MD, DrPH, MBA; Douglas J. Perkins, PhD; Christophe G. Lambert, PhD

1Vanderbilt University Medical Center, Nashville; 2University of New Mexico, Albuquerque, NM; 3VA Tennessee Valley Healthcare System, Nashville, TN; 4VA New Mexico Healthcare System, Albuquerque, NM

Introduction
Incomplete coding of phenotypes in electronic health record (EHR) data poses a major obstacle to generating high-quality data for clinical care and research.1 Post-traumatic stress disorder (PTSD) and self-harm may be documented with diagnostic codes in only 2-19% of cases.2,3 Undocumented conditions impede timely screening, diagnostics, referrals and treatment. Further, lack of coding complicates our understanding of health disparities, especially if accuracy varies across patient groups. In this study, we characterize disparities in PTSD and self-harm among US veterans using coded and imputed phenotypes, exploring how documentation practices impact our understanding.

Methods
We accessed US Department of Veterans Affairs national EHR patient data for 2000-2020. Positive unlabeled learning with XGBoost estimated the probability of PTSD and self-harm for a random sample of veterans born after 1940 and receiving care in the VA system for ≥2 years. We estimated the impact of sociodemographic factors on risk of PTSD and self-harm using coded and imputed phenotypes, as well as on the risk of being a probable but uncoded (i.e., missed) case vs a coded case. Logistic regression models estimated associations with race/ethnicity, decade of birth, sex, homelessness, unemployment, county urbanization, and county socioeconomic deprivation covariates. These models also controlled for major mental health conditions known to be associated with self-harm and PTSD.

Results
Our study included 1,022,102 veterans. PTSD was coded for 24.0%. Imputation identified few patients at high risk without a coded diagnosis, adding only 355 cases. The odds of PTSD were 24%-66% higher among minority veterans compared to Non-Hispanic White (NHW) veterans. Female veterans were slightly less likely than their male counterparts to have PTSD documented in the EHR and unemployment conferred 21% increased risk of PTSD.

Self-harm was coded in 1.9% of the population and imputation identified 25,401 additional patients, raising the self-harm rate to 4.3%. Non-Hispanic Black (NHB) veterans were at 25% increased odds of probable but uncoded self-harm compared to NHW veterans. However, the odds of self-harm were 51% higher among NHW veterans. Odds ratios for self-harm increased from 1.96 [1.89-2.04] to 2.48 [2.42-2.55] for homelessness and from 2.06 [1.98-2.14] to 2.68 [2.61-2.76] for unemployment when modeling with coded and imputed cases, respectively. For other sociodemographic features, imputation did not significantly impact our understanding of health disparities.

Conclusion
Among US veterans, we documented disparities in PTSD and self-harm associated with age, race, sex, and economic indicators. Imputation identified few additional cases of PTSD and did not impact our understanding of disparities. On the other hand, imputation revealed many cases of self-harm and we found analyses relying on coded self-harm significantly underestimated associations with homelessness and unemployment. Correcting disparities analyses with enriched EHR-phenotypes through imputation may provide important insight into screening/diagnostic biases, populations in need of additional resources, and the types of interventions necessary to improve veteran health.

References
An AI Model to Screen for SARS-CoV-2 in Emergency Department Clinical Notes

John Del Gaizo1, Garett Hayes2, Jihad S Obeid1, MD PhD
1Biomedical Informatics Center, Medical University of South Carolina, Charleston, SC, USA; 2College of Medicine, Medical University of South Carolina, Charleston, SC, USA

Introduction

We present an AI model for the rapid screening of potential COVID-19 (+) patients upon presentation to the ED (emergency department), specifically from physician-authored ED notes. Such a model can provide public health benefits through (1) rapid testing; (2) real-time disease incidence metrics; and (3) automated identification of key symptoms through X-AI (eXplainable Artificial Intelligence) algorithms such as Grad-CAM (Gradient – Class Activation Mapping).

Methods

We extracted 9,597 physician-authored ED notes (2,588 positive, 7,009 negative) from the MUSC RDW (Medical University of South Carolina Research Data Warehouse) from 2 groups (IRB Pro00100990): (1) 2,588 notes from patients with COVID (+) lab tests within 2 weeks of the note; and (2) a demographically-matched control set of 7,009 notes from subjects with negative lab tests and without coronavirus ICD codes. The notes were recorded between 03-01-2020 and 01-14-2021. The data was split on September 15 into 2 groups based on the occurrence of 2 major COVID incidence waves within the time frame.

Table 1. Dataset sizes for AI models, split on 09-15-2020 in occurrence with COVID incidence rates.

<table>
<thead>
<tr>
<th>Time Frame</th>
<th>Train Sample Size</th>
<th>Hold Out Sample Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wave 1 03-01-2020 to 09-15-2020</td>
<td>3,593</td>
<td>1,542</td>
</tr>
<tr>
<td>Wave 2 09-16-2020 to 01-14-2021</td>
<td>3,124</td>
<td>1,338</td>
</tr>
<tr>
<td>Total 03-01-2020 to 01-14-2021</td>
<td>6,717</td>
<td>2,880</td>
</tr>
</tbody>
</table>

Our data processing pipeline consisted of the following steps: (1) regular expressions fine-tuned for medical EHR notes, (2) tokenization, (3) numericizing, (4) word embedding, and (5) a CNN (convolutional neural network) for COVID prediction. We next implemented code that applies Grad-CAM to a text input, and outputs HTML (Figure 1). The rendering enlarges and reddens tokens in accordance with prediction importance.

Results

The hold out set AUC values were .82 and .77 on the models trained on the Wave 1 and Wave 2 data, respectively.

Figure 1. Grad-CAM algorithm. PHI has been blotted out, underscore (0_0…) words are simply padded elements.

Conclusion

The AI model was comparable to the AUC derived from EM doctors2. However, the model AUC decreased over time. One possibility is that increased familiarity of ED physicians with standard COVID presentations led to documentations changes.

We are in the early stages of experimentation with other architectures, such as regression models applied to BERT embeddings3. Of particular interest for X-AI is the HAN (hierarchical attention network)4, which calculates attention weights that are essentially discrete probability distributions for both tokens within a sentence, and sentences within a sample.

References

Leveraging Ontologies for the Aggregation of Normalized Pharmacy Content
Michael Denton MS, RN, Tosh Kartchner RN, Jonathan Herbert RN, Dean Woolstenhulme MS
3M Health Information Systems, Murray, Utah

Introduction
Multiple variations exist between local and standard pharmacy terminologies, limiting interoperability and the creation of accurate cross terminology datasets (map sets, value sets). By leveraging a relational terminology server with a controlled medical vocabulary, we store, translate, and aggregate both consumer and standard pharmacy terminologies in a unified structure, enabling flexible tooling to create standardized datasets for multiple use cases and in varying structures. The goal of this project is to improve the maintenance and usability of cross-terminology pharmacy content, providing end users with comprehensive coverage and multiple levels of granularity for a given use case.

Methods
The project consists of two phases: translation and attribution. In the translation phase, subject matter experts (SMEs) evaluate standard and consumer terminologies to convert to a normalized structure that can be stored in our relational terminology server. During the translation process, concepts are defined with sets of definitional attributes and where gaps exist, attributes can be newly created. The attribute sets are grouped together in an ontological structure which can be organized and aggregated based on type and granularity, resulting in normalized groups of “like” attributes used to define clinical drugs across terminologies with different levels of specificity, increasing the usability of pharmacy content. In addition to making content more flexible, the structure also improves the ability to maintain datasets, integrate standard updates and consumer requests seamlessly while maintaining concept permanence, and allow content to evolve gracefully once a formal definition is defined. During the attribution phase, SMEs map sets of definitional attributes and relational operators necessary for defining group(s) of content, narrowing clinical drugs to the set of concepts with the level of specificity desired for a given use case. This is done by leveraging the attribute ontology and the relationships between each attribute type. For this project, a “clinical drug” is defined as a concept attributed with an ingredient and one or more definitional attributes: route, form, or packaging. Optional attributes that could be included, but not required to define a clinical drug, are the ingredient strength, brand name, and modifiers. For example, if we wanted to aggregate clinical drugs for “intravenous lidocaine”, we would attribute the ingredient “lidocaine hydrochloride” and additional attributes such as a route (infusion, injection, intravenous) or a form (liquid, solution, suspension). Once attributed, the tooling will aggregate clinical drugs using the forementioned attribute types and any “like” attributes depending on the relational operators. SMEs will continue to evaluate and revise the set of definitional attributes until the correct set of clinical drugs are captured for a given use case.

Results and Discussion
The project has resulted in the successful translation of 32 source terminologies, including First Databank (FDB), Gold Standard Drug Database (GSDD), National Drug Code (NDC) Directory, National Drug File – Reference Terminology (NDF-RT), RxNorm, SNOMED CT, Vaccines Administered (CVX), and other consumer source terminologies. This provides us with 323,636 distinct clinical drug combinations which can be leveraged in our attribution process. The combinations utilize a distinct set of attributes including 37,844 ingredients, 129 routes, 884 forms, 1,148 packages, 76,326 ingredient strengths, 50,603 brand names, and 307 modifiers. We also leverage pharmacological classes from standard terminologies, enabling us to aggregate clinical drugs from 156 hierarchical classes and 1,542 therapeutic classes. This allows us to evaluate attributes from the clinical drugs in a desired class which can be leveraged to include cross-terminology concepts with the same definitional attributes. In addition to the drug hierarchies from standard terminologies, The Drug Ontology (DrOn) was also assessed for feasibility of incorporation into our relational terminology server. DrOn was useful as a reference source, but limitations exist for translating standard and local terminologies beyond RxNorm and would require additional tooling for our use case. In addition, we needed a way to maintain multiple terminology sources and cross-terminology datasets to reduce redundancies and prevent semantic drift, which is beyond the scope of DrOn.

Conclusion
By using a relational terminology server with a controlled medical vocabulary, we are able to store, translate, and aggregate pharmacy content accurately in a normalized structure. Then, additional tooling was created to build accurate cross-terminology data sets that are flexible and maintainable for varied use cases.

References
Assessing the Use of HL7® FHIR® Among Healthcare Apps

Brian J. Douthit, MSN, RN-BC1, Guilherme Del Fiol, MD, PhD2, Chloe Canon, BM3, Jessica Branski, BS3, Titus Schleyer, DMD, PhD4,5, Rachel L. Richesson, PhD, MPH3
1Duke University, Durham, NC; University of Utah, Salt Lake City, UT; 3University of Michigan, Ann Arbor, MI; 4Regenstrief Institute and 5Indiana University School of Medicine, Indianapolis, IN

Introduction

The Substitutable Medical Applications, Reusable Technologies (SMART) on Fast Healthcare Interoperability Resources (FHIR) is an increasingly popular standard on which to develop and deploy interoperable health apps. The FHIR standard has enabled greater access to electronic health record (EHR) data, but plug-and-play interoperability is still limited. However, FHIR is continuously being modified to maximize interoperability, such as through the development of US Core profiles. This work serves to inform priority areas for FHIR profile development by assessing the data requirements among a representative sample of SMART on FHIR apps and the FHIR resources that were used to meet those requirements.

Methods

FHIR apps were sampled from two sources: the SMART App Gallery (https://apps.smarthealthit.org/) and the yearly AMIA/HL7 FHIR Applications Competition. This was done to ensure the apps 1) were of high quality, 2) included varying purposes and clinical foci, and 3) could be used with multiple EHRs. To gather information regarding resource use, app authors were contacted, and open-source code was analyzed when available. If neither option was successful, app demos were analyzed for data requirements which were subsequently mapped to the appropriate FHIR resource.

Results

Of the 90 apps available on the SMART gallery and the 27 apps submitted to the AMIA FHIR app competition between 2018 and 2020, 51 apps were included in the analysis. Apps were excluded from analysis if resources could not be acquired from open-source coding, app demos, or from contacting the app authors. Table 1 summarizes the findings by reporting the raw count and frequency of FHIR resource use among the sample of apps. Condition was the most frequently used non-normative resource, which may be a priority area for development. MedicationStatement, FamilyMemberHistory, Questionnaire, and QuestionnaireResponse were among the top 12 most used resources without associated US Core profiles.

Table 1. Twelve most frequent FHIR resources used in app sample

<table>
<thead>
<tr>
<th>Resource</th>
<th>FHIR Maturity Level</th>
<th>Number of apps</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient*</td>
<td>Normative</td>
<td>41</td>
<td>80.4%</td>
</tr>
<tr>
<td>Observation**</td>
<td>Normative</td>
<td>39</td>
<td>76.5%</td>
</tr>
<tr>
<td>Condition*</td>
<td>3</td>
<td>22</td>
<td>43.1%</td>
</tr>
<tr>
<td>Medication*</td>
<td>3</td>
<td>17</td>
<td>33.3%</td>
</tr>
<tr>
<td>AllergyIntolerance*</td>
<td>3</td>
<td>15</td>
<td>29.4%</td>
</tr>
<tr>
<td>Procedure*</td>
<td>3</td>
<td>12</td>
<td>23.5%</td>
</tr>
<tr>
<td>DiagnosticReport* (ResultsSection)</td>
<td>3</td>
<td>10</td>
<td>19.6%</td>
</tr>
<tr>
<td>MedicationStatement</td>
<td>3</td>
<td>8</td>
<td>15.7%</td>
</tr>
<tr>
<td>FamilyMemberHistory</td>
<td>2</td>
<td>7</td>
<td>13.7%</td>
</tr>
<tr>
<td>Questionnaire</td>
<td>3</td>
<td>5</td>
<td>9.8%</td>
</tr>
<tr>
<td>QuestionnaireResponse</td>
<td>3</td>
<td>5</td>
<td>9.8%</td>
</tr>
<tr>
<td>CarePlan*</td>
<td>2</td>
<td>5</td>
<td>9.8%</td>
</tr>
</tbody>
</table>

*Has an associated US Core Profile; **Specific observations (e.g., vital signs) have an associated US Core Profile

Conclusion

Continual assessment of FHIR resource use could help to inform the development of the standard and facilitate the development of US Core profiles. Additional sources of FHIR apps should be considered in future iterations of this work, and app authors should be encouraged to share their FHIR resource usage so analyses such as this may be done more readily.
Development of a Drug Allergy Alert Tiering Algorithm for Penicillins and Cephalosporins

Heba Edrees, PharmD1,2, Diane L Seger, RPh3, Ying-Chih Lo, PhD1, David W Bates, MD, MSc1,4, Li Zhou, MD, PhD1,4

1Brigham and Women’s Hospital, Boston, MA; 2MCPHS University, Boston, MA; 3Mass General Brigham, Somerville, MA; 4Harvard Medical School, Boston, MA

Introduction

Clinical decision support (CDS) has become a useful tool in reducing adverse drug events, including drug allergy interactions. Drug allergy alerts are a common CDS problem because alerts are triggered strictly based on the medication being ordered and the patient’s documented allergy. This results in too many allergy alerts being generated and has resulted in an increase in drug allergy alert overrides. Patient allergy alerts are overridden about 80% of the time and about 96% of them are overridden appropriately. In our institution, antibiotics, specifically penicillins and cephalosporins, were the second highest drug class with allergy alerts overrides (14%), following opioids (30%). Current drug allergy alerting mechanisms can be improved, as they negatively impact physicians’ time and patient safety due to alert fatigue. The objective of this study is to determine the frequency of overrides for penicillin and cephalosporin drug allergy alerts and create a rule-based algorithm to reduce the number of inappropriate drug allergy alerts.

Methods

Drug allergy alert and override data was reviewed for penicillins and cephalosporins, over a one-year period, for patients at a large Massachusetts Healthcare System. The data will be used to develop a rule-based algorithm to classify drug allergy alerts into three groups: no alerts presented to the user, non-interruptive (informational) alerts, and interruptive alerts that require a coded response or action. The rule-based algorithm will include factors such as drug class or cross-sensitivity matches, reaction types with the designated severities (categorized as high, medium, or low), patient comorbidities, and allergy alert history. To date, patients’ drug allergies and reactions, medications that triggered the drug allergy alert, and overrides were reviewed.

Results

Drug allergy alerts for penicillin and cephalosporins were overridden 61.4% of the time. Of the drug allergy alert overrides, 15% were for medications in the same drug class and 85% were for drugs that had a potential for cross sensitivity. Of all the reactions documented, 7% of the overridden alerts had a reaction classified as severe, 20% were medium severity, 38% were low severity, and 35% were unknown severity because a reaction was not documented. The most common high severity reactions were anaphylaxis, angioedema, and shortness of breath. The most common low severity reactions were rash, gastrointestinal upset, and itching. Of the drug allergy alerts with a cross sensitivity match, 28% were for patients with a documented penicillin allergy and a medication order for a first-generation cephalosporin, with a 41% low severity reaction. In addition, 30% of overridden drug allergy alerts were for patients with a penicillin allergy and an order for a third-generation cephalosporin, with a 36% low severity reaction.

Conclusion

Providers frequently overrode drug allergy alerts when the allergy reaction severity was classified as low and the alert was for medications that had a potential for cross sensitivity. This preliminary data shows that the number of drug allergy alerts, shown to the provider, could potentially be reduced by filtering by reaction severity and risk of cross-sensitivity. This data is being used to develop an algorithm to redesign drug allergy alerts that will reduce override rates, minimize clinician interruption, and improve patient safety.

References

Roua El Kalach, PharmD, BCPS1; Lori Moore, PharmD, CAPT2; Ulrica Andujar, MPH3

1.2 National Center for Immunization and Respiratory Diseases (NCIRD), Centers for Disease Control and Prevention (CDC), Atlanta, GA; 2Chicкахew Health Consulting, Atlanta, Georgia

Abstract Summary
This presentation discusses the key components for the successful development and implementation of a federal external information system (ExIS) to interface with the Centers for Disease Control and Prevention (CDC) Vaccine Tracking System (VTrckS), and provides an overview of implementing the technology for ordering COVID-19 vaccines during the pandemic response. This report also highlights opportunities to leverage technologies to enhance the supply chain distribution.

Introduction
The Centers for Disease Control and Prevention (CDC) uses the Vaccine Tracking System (VTrckS), a secure, web-based information technology system, to support ordering and distribution of publicly funded routine vaccines. VTrckS manages and distributes within territories, states, and local health departments, as well as federal entities. The Vaccine Ordering System (VOP), described in this presentation, supports the ordering and distribution of COVID-19 vaccines through the Federal Retail Pharmacy Program, the health center COVID vaccine program, and the Rural Health Clinic COVID Vaccine Distribution Program. Further, following careful consideration and assessment, a federal external information system (federal ExIS) was developed to allow 30 commercial and 7 federal agency partners with more than 40,000 provider locations (pharmacies, dialysis clinics, mobile mass vaccination clinics, and federal clinics) to place and receive COVID-19 vaccine orders. This presentation highlights the key factors leading to a successful development and implementation of a federal ExIS, and opportunities for leveraging technologies to expand federal supply chain capacity.

Methods
CDC reviewed past and current ExIS awardee activities utilized for the routine vaccine program as well leveraging tools and resources developing CDC’s Immunization Services Division Immunization Information Systems Support Branch to support ExIS implementations over the past few years. Examples of tools and resources included the CDC-developed ExIS File Specifications, Processing VTrckS Orders document, and the ExIS Usability Best Practices Catalog. CDC also leveraged subject matter experts (SMEs) actively supporting and providing technical assistance to guide awardees through ExIS implementation and enhancements. By compiling existing resources, tools, and SME expertise, CDC was able to rapidly develop a team to support the planning, analysis, design, implementation, testing, evaluation, and support of a federal pandemic ExIS system called the VTrckS Partner Ordering Portal (VPOP).

Principal Findings
VPOP has been adopted by more than 30 commercial and federal partners for placing COVID vaccine orders and has contributed to the distribution of approximately 1/3 of the total distributed US doses. The primary requirement for successful VPOP implementation was its ability to interface with VTrckS while supporting multiple data sources in order to rapidly expand ordering capacity to a larger provider base to ensure rapid distribution of COVID-19 vaccines.

Table 1. Summary of key drivers for the successful and timely development and deployment of VPOP.

<table>
<thead>
<tr>
<th>Key Factors</th>
<th>Description</th>
<th>Applicability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard Specifications Guidelines</td>
<td>ExIS File Specifications document outlined the required data elements and business rules that had to be implemented to interface with CDC VTrckS.</td>
<td>The guide contains file specifications for inbound and outbound interfaces for integration of VTrckS and ExIS systems. The specifications document served as the data dictionary for developing the ExIS and ensuring successful information exchange with VTrckS (i.e., interoperability).</td>
</tr>
<tr>
<td>Usability Best Practices Catalog</td>
<td>Outlined best practices for vaccine ordering and inventory tracking, and described business rules that minimized errors, supported accountability, and were efficient (e.g., in terms of the number of screens, steps, people involved, and time to complete tasks).</td>
<td>The Usability Best Practices Catalog was inspired from design observations of awardee-training materials, ExIS functionality demonstrations, focused workshops, the Modeling of Immunizations Registry Workgroup, IIS Inventory Management Operations Guide, and other published material.</td>
</tr>
<tr>
<td>Agile Development Methodology</td>
<td>The development team adopted the time-boxed project cycles known as sprints to release new application features.</td>
<td>The agile development methodology allowed the team to adapt easily and quickly to change, release features at predictable delivery dates, engage stakeholders and users in testing, and integrate changes to achieve better user satisfaction with every sprint.</td>
</tr>
<tr>
<td>User Acceptance Testing/Test-Driven Development</td>
<td>Application owners worked closely with users in the last environment to gain a clear understanding of business and workflow requirements to meet user needs before every release.</td>
<td>The User Acceptance Testing environment was established, and access was granted to the CDC testing team and to a group of users from several partners interested in testing application features and providing feedback before every release.</td>
</tr>
<tr>
<td>Cloud-Based Development</td>
<td>Cloud concepts and technologies provided ample resources to support software productivity through code development, testing, and code repositories.</td>
<td>Cloud hosting offered predictable server uptime, security, scalability, back-up and business continuity, and location flexibility.</td>
</tr>
<tr>
<td>Security</td>
<td>Cloud service providers protected secured and encrypted solutions, backup recovery, firewalls, identity management, and data isolation and storage segregation.</td>
<td>The application leverages: - Secure cloud-based solution - Authorized access only to authorized users through password protection - Two-factor authentication is required</td>
</tr>
<tr>
<td>User-friendly Interface</td>
<td>Intuitive, easy to access and use application.</td>
<td>The application design offered an one-stop shop for placing COVID vaccine orders or performing other relevant activities on one screen. Most business processes could be completed with only a few clicks.</td>
</tr>
<tr>
<td>Flexible Implementation and Provision</td>
<td>Enrolment and credentialing of providers.</td>
<td>The application design adopted a “hub and spoke” model allowing partners to manage and authorize access to their users. Authorized users are easily granted access to the system once they are added as a contact to a location.</td>
</tr>
<tr>
<td>Stakeholder and Partner Involvement</td>
<td>Partner and stakeholder engagement was conducted throughout project life cycle.</td>
<td>Partners and stakeholders have been motivated to serve and be a part of the project targeted at serving public health during the pandemic.</td>
</tr>
<tr>
<td>Crowdsourcing Platform</td>
<td>Crowdsourcing is a way to collect user’s knowledge or feedback.</td>
<td>The application hosted a crowdsourcing platform for all users to share feedback “on the fly” or suggest features through the feedback or project management features while completing regular ordering or processing activities in the portal. Crowdsourcing tools also provide transparency to requesting end users by allowing them to track/#/enhancement was approved, entered development, and which sprint cycle it would be released in.</td>
</tr>
<tr>
<td>User Guide/Training</td>
<td>User-centric manual providing instructions on the use of the application.</td>
<td>The portal enables users to access the user guide through the Help feature and walk through the steps required to complete an activity. Adopting the “hub and spoke” model saved a lot of training effort and allowed us to meet training needs through the “train the trainer” approach.</td>
</tr>
</tbody>
</table>

Conclusion
By leveraging existing public health system implementations, CDC and partners demonstrated the additional potential in leveraging technologies to expand the supply chain and distribution process. Additional areas of improvement have also been identified including the automation of ExIS systems. CDC plans to continue this initiative to expand across all ExIS systems ordering vaccine to enhance processes and modernize public health data.

References
Race and Ethnicity Agreement Between Medicaid, Health Information Exchange, and Electronic Health Records

Sarah El-Azab¹, MS, Siu Hui², PhD, Heidi Hosler², BS, Katie S. Allen², BS
University of Michigan, Ann Arbor, Michigan¹, Regenstrief Institute, Indianapolis, Indiana²

Introduction
High quality race and ethnicity data is essential for research that assesses health disparities between populations¹. However, in the observational data sources most frequently utilized by researchers to do so, the quality of these data elements can be subpar. In 2018, the Centers for Medicare and Medicaid Services (CMS) reported that the quality of the Medicaid race and ethnicity data for 20 states was “highly concerning” or unusable². Electronic health records (EHRs) suffer from similar data quality issues, due to the challenges of collecting race and ethnicity in clinical settings¹. These issues are likely to be compounded in health information exchanges (HIEs), which collect data from numerous EHRs. Given these data quality concerns, it is critical for researchers to understand the strengths and weaknesses of each data source or risk unknowingly weakening the validity of their research. This study compares race and ethnicity data between a state HIE, Medicaid, and the EHRs of an academic medical center (AMC) and a federally qualified health center (FQHC) for the same cohort of patients.

Methods
A cohort of adult Medicaid enrollees was linked to a HIE, AMC, and FQHC by using exact matching on demographic data points. Patient identifiers, race, and ethnicity data were extracted from all four data sources. While the AMC and FQHC contribute to the HIE, race and ethnicity data was extracted directly from their respective EHRs. All available race data was categorized as “White”, “Black or African American”, “Asian or Pacific Islander”, “American Indian or Alaska Native”, “Multiracial”, “Other”, or “Unknown”. Ethnicity was similarly categorized as “Hispanic or Latinx”, “Not Hispanic or Latinx”, and “Unknown”. For each data source, race or ethnicity was assigned by taking the mode of all available data per patient. In the case that a patient was multimodal, they were classified as “Unknown.” Pairwise agreement between the data sources was assessed using percent agreement and Cohen’s Kappa statistic.

Results
There was wide variation in the race and ethnicity categories among the data sources. Medicaid did not collect data on Multiracial patients and the two EHRs did not collect data for “Other.” There were also considerable differences in the percentage of White and Black patients across the data sources. For the same cohort, Medicaid had a higher percentage of White patients (70.46%) than the HIE (38.66%), the AMC (52.44%), and the FQHC (44.43%). Medicaid also had a lower percentage of Black patients (20.04%) than the HIE (32.67%), the AMC (38.63%), and the FQHC (38.64%). The pairwise agreements between the data sets are shown in Table 1.

<table>
<thead>
<tr>
<th>Data Source Pair</th>
<th>Race: Percent Agreement</th>
<th>Race: Cohen’s Kappa</th>
<th>Ethnicity: Percent Agreement</th>
<th>Ethnicity: Cohen’s Kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medicaid – HIE</td>
<td>36.2%</td>
<td>0.005</td>
<td>55.3%</td>
<td>0.006</td>
</tr>
<tr>
<td>Medicaid – AMC</td>
<td>45.5%</td>
<td>0.003</td>
<td>62.9%</td>
<td>0.008</td>
</tr>
<tr>
<td>Medicaid – FQHC</td>
<td>40.9%</td>
<td>0.011</td>
<td>37.4%</td>
<td>0.001</td>
</tr>
<tr>
<td>HIE-AMC</td>
<td>75.7%</td>
<td>0.627</td>
<td>71.0%</td>
<td>0.375</td>
</tr>
<tr>
<td>HIE- FQHC</td>
<td>71.9%</td>
<td>0.576</td>
<td>55.5%</td>
<td>0.265</td>
</tr>
<tr>
<td>AMC - FQHC</td>
<td>79.1%</td>
<td>0.655</td>
<td>43.0%</td>
<td>0.077</td>
</tr>
</tbody>
</table>

Table 1. Pair-wise comparison across data sources with percent agreement and Cohen’s Kappa statistic.

Conclusion
The AMC and FQHC had substantial pairwise agreement with the HIE, likely because they contribute data to it. However, there is poor agreement between Medicaid and the EHR-based data sources. Further research is needed to assess how race and ethnicity is collected in these settings and what is contributing to these differences.

References
Usability and Accuracy Evaluation of the Fitbit Versa 3 Smart Watch Compared to Research-Grade Devices

Chuka Emezue, PhD, MPH, MPA\textsuperscript{1}, Chelsea Howland, MSN\textsuperscript{1}, Malaika R. Gallimore, RN, MPH\textsuperscript{1}, Katrina Boles, MS\textsuperscript{1}, Amy Grimsley, MSN, RN, CCRN-K\textsuperscript{1}, LeeAnne B. Sherwin, PhD, MS, FNP-BC\textsuperscript{1}, Allison Anbari, PhD, RN\textsuperscript{1}, Blaine Reeder, PhD\textsuperscript{1}, Jo-Ana D. Chase, PhD, APRN-BC\textsuperscript{1}

\textsuperscript{1}University of Missouri, Columbia, MO, USA.

Introduction

Consumer-graded wearable devices continue to enjoy increasing consumer adoption and acceptance. However, studies show a tendency for wearables to underestimate steps in controlled testing and overestimate steps in free-living settings. In addition, rapid technological changes and short product release cycles surpass device testing efforts by skilled evaluators. At the University of Missouri Precision START laboratory, we devised a strategy to inform research and health care decisions using a multi-step methodology involving within-team, lab- and field-based tests to rapidly evaluate the Fitbit Versa 3 for accuracy, usability, and feature availability\textsuperscript{1}. This study reports within-team evaluations of the Fitbit Versa 3 compared to two research-grade control devices.

Methods

In February 2021, a five-person team completed lab-based testing of the Fitbit Versa 3 (Fitbit) smart watch, a research-grade control device - the hip-worn ActiGraph wGT3X-BT (GT3X) - and wrist-worn GT9X-Link (GT9X). Evaluators wore the Fitbit and GT9X on their non-dominant arm and the GT3X control on their dominant hip. During 7 days of continuous wear, participants recorded free-living activities on a weekday (Day 3) and a weekend (Day 6). Scripted tasks for walking and stair-climbing activities were completed on Day 4. For walking activity, participants walked a straight pathway with a 0% incline at three self-paced speeds for an 8-minute each for “very light walking,” “light walking,” and “moderate walking,” followed by 2-minute rest periods. Evaluators completed a series of post-test checklists from our stepwise methodology\textsuperscript{1}, including the 10-item system usability scale (SUS), 12-item user experience checklist and a feature availability survey. The SUS was applied separately to the Fitbit smart watch and mobile app.

Results

We calculated the median absolute percentage error (or MdAPE= median (p1,p2,⋯,pN), where p represents each participant). MdAPE is an outlier- resistant measure and reduces bias in favor of lower scores, thus advantageous over the mean absolute percentage error (MAPE). Overall, the Fitbit vs. hip-worn GT3X control achieved the closest agreement with % error of 54% while the Fitbit vs. GT9X comparison had the widest % error of 94.9% (i.e., least precision). Stair Climbing Activity: The % error for total stair climbing activity between the Fitbit vs. hip-worn GT3X control was 20.5%, and for the Fitbit vs. wrist-worn GT9X, % error was 40.85%; indicating better agreement between the Fitbit and GT3X control. Walking Activity: For self-rated exertion, moderate walking had the lowest % error. With the Fitbit vs. GT3X control, % error was 91% for “very light walking,” 71.1% for “light walking,” and 16.2% for “moderate walking.” For the Fitbit vs. wrist-worn GT9X comparison, % error was 95.2% for “very light walking,” 94.9% for “light walking,” and 93.9% for “moderate walking.” This difference in device accuracy held true for Daily-Living Activity Step Counts: % error for the Fitbit vs. GT3X control was 41.45%, while the Fitbit vs. GT9X % error was 44.55%. The median absolute percentage error was lower in daily living activities. Usability: Average SUS scores for all five evaluators on the Fitbit Watch was 87.0 (82.5, 90, 82.5, 100, 80) or excellent usability, and for the Fitbit Mobile App, the usability score was 92.5 (87.5, 92.5, 90, 97.5, 95), indicating excellent usability.

Conclusion

Our results indicate that the Fitbit Versa 3 smart watch did not accurately measure step counts compared to the hip-worn ActiGraph GT3X in controlled conditions. While SUS results indicated excellent usability for watch and app, device error was higher in some activities (e.g., during “moderate walking”), raising concerns about using the Fitbit alone for research in free-living environments. Differences in home arrangement may be associated with reduced Fitbit accuracy even with controlled tests using zero inclines, defined distances, and explicit test instructions. In addition, the Actigraph wrist-worn GT9X appears to be a less accurate comparison device than the GT3X control device. Next steps will evaluate the Fitbit Versa 3 in comparison in participant field tests.

References

Observations on Documentation of Alcohol Use in Real-World Data

Emilia Farcas, PhD¹, Michael Hogarth, MD², Alison Moore, MD, MPH³
¹Qualcomm Institute, ²Division of Biomedical Informatics, ³Division of Geriatrics, Gerontology & Palliative Care, University of California San Diego, La Jolla, CA

Introduction
Excessive alcohol use is a leading cause of preventable death in the US, with an average of 95,000 deaths annually. During the pandemic, excessive alcohol use and its harms increased. While screening for excessive alcohol use has many barriers, the rapid adoption of Electronic Health Record (EHR) systems in the US over the past decade has enabled implementation of electronic versions of standardized questions and screening tools to assess for excessive alcohol use. To assess the utility of EHR data for screening and assessment of excessive alcohol use, we examined alcohol use items available in one health system’s EHR over a five-year period. Though one study¹ of 500 records examined alcohol use documentation, it had a different purpose (i.e., reasons for use of free text).

Methods
We conducted a retrospective analysis using deidentified EHRs from the UC San Diego Health Clinical Data Warehouse. The cohort included 486,290 patients 18 years or older with 5,614,732 encounters between 11/2014 – 11/2019. Alcohol data were found in structured fields in three separate data structures: (1) social history, including alcohol use (i.e., yes, no, not currently, never) and the first three items from the Alcohol Use Disorders Identification Test (AUDIT-C): frequency of use (i.e., never, monthly of less, 2-4 times a month, 2-3 times a week, 4 or more times per week), number of drinks per day (i.e., 1-2, 3-4, 5-6, 7-9, 10 or more), frequency of drinking 6+ drinks on an occasion, and comments; (2) the complete 10-item AUDIT; and (3) Alcohol/week, including quantity and type of beverage (i.e., glasses of wine, cans of beer, shots of liquor). Besides paragraphs of free text, some clinical notes (e.g., progress notes, patient instructions) also contain a section that starts with “Alcohol use:” followed by one or more alcohol items, allowing pattern matching. We divided the cohort into two groups based on whether they had their primary care assigned to UC San Diego (i.e., without vs with). We analyzed all encounters and also by department (e.g., emergency and primary care). All encounters include 857,744 encounters with no department assigned.

Results
Most alcohol documentation occurred via the AUDIT-C and clinical notes. Alcohol data were stored in different data fields at different encounters. The AUDIT is collected in primary care departments, whereas the Alcohol/week data are collected in the emergency departments. Compared to patients without primary care assigned to UC San Diego, a larger proportion of patients with primary care assigned had alcohol data documented, even for emergency encounters.

<table>
<thead>
<tr>
<th>Patients with alcohol data recorded at any encounter of the specified type</th>
<th>1. AUDIT-C from Social History</th>
<th>2. AUDIT</th>
<th>3. Alcohol/week</th>
<th>Any source 1-3</th>
<th>4. Clinical Notes</th>
<th>Any source 1-4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patients without primary care assigned to UC San Diego</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All encounters</td>
<td>3,389,434</td>
<td>401,821</td>
<td>94,263 (23%)</td>
<td>1,170 (0.3%)</td>
<td>10,290 (3%)</td>
<td>98,417 (24%)</td>
</tr>
<tr>
<td>Emergency</td>
<td>134,869</td>
<td>75,673</td>
<td>2,342 (3%)</td>
<td>-</td>
<td>10,290 (14%)</td>
<td>12,333 (16%)</td>
</tr>
<tr>
<td>Patients with primary care assigned to UC San Diego</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All encounters</td>
<td>2,225,298</td>
<td>84,469</td>
<td>58,783 (70%)</td>
<td>12,247 (14%)</td>
<td>5,456 (6%)</td>
<td>62,191 (74%)</td>
</tr>
<tr>
<td>Emergency</td>
<td>34,148</td>
<td>18,269</td>
<td>1,805 (10%)</td>
<td>-</td>
<td>5,456 (30%)</td>
<td>6,992 (38%)</td>
</tr>
</tbody>
</table>

Table 1. The availability of alcohol data in various fields in the EHR

Discussion
Alcohol documentation occurred most frequently in primary care as part of standardized screening questions. Clinical notes were also a common source of alcohol data. In future work, we plan to use natural language processing on free text comments and clinical notes to identify additional alcohol data and analyze inconsistencies between these and the structured fields. While the EHR is a promising means of gathering data on alcohol use and identifying excessive alcohol use, we found that alcohol use data is recorded variably in the EHR and fragmented across multiple encounters. Standardized approaches are needed to reliably gather alcohol use information to drive personalized assessments.

Acknowledgements
This work has been supported in part by NIH award 5R56AG067393 and NIH Grant UL1TR001442 of CTSA funding.

References
Implications of Clinical Bias When Building Electronic Algorithms for Medical Record Phenotyping: Lessons from Type 2 Diabetes Mellitus

Annika Faucon, BS¹, Megan Shuey, PhD¹, Lea K. Davis, PhD¹, Nancy Cox, PhD¹

¹Vanderbilt University, Nashville, TN, USA

Abstract

The advent of electronic health record systems has made clinical phenotyping paramount. The process of delineating case and control populations, can introduce unexpected bias due to how, why, and from where the data was ascertained. Striking a balance between sensitivity, specificity, and introduced bias is essential for use in genetic and epidemiologic studies. Using three algorithms for type two diabetes mellitus (T2DM), we highlight clinical differences and potential confounding bias due to the modifications made for capture improvement.

Introduction

Genetic and epidemiologic studies require high specificity and large sample sizes to accurately identify correlations. Diabetes, for example, is a complex condition defined clinically by a hemoglobin A1C (HbA1C) value above 6.5%, a fasting blood sugar above 125 mg/dL, or an oral glucose tolerance test two-hour blood sugar above 200 mg/dL. An existing algorithm¹ for T2DM deployed across six eMERGE EHRs successfully identified patients with T2D, but application of the algorithm to present day EHRs leaves a large portion of patients with diagnosed T2DM unclassified (53.9%); limiting statistical power in downstream analyses. In addition to updating the existing algorithm, we evaluated the decision to define controls by their HbA1c or glucose measures. Throughout this process, we evaluated population level changes to characterize the biases introduced by definition change.

Results, Discussion, and Conclusion

The previous algorithm classified case values based on ICD codes, laboratory values, and timing of medications (e.g. insulin should not precede T2DM-specific drugs, like metformin). However, new medications for T2DM have become available in the past decade since eMERGE first published. Because medical science continues to evolve over time, updating useful tools in a sensible fashion is critical to continued building of knowledge.

Over 14% of Americans are estimated to have diabetes², and despite individuals with diabetes being more likely to utilize the EHR, phenotyping strategies fail to capture a similar disease frequency. Using BioVU, a Biobank-Linked EHR, we observed simply updating the algorithm to include ICD10 clinical modification codes results in an improvement to recall such that the prevalence of T2DM is 10.3%. Including newer medications further improved capture by increasing prevalence to 12.6% in individuals of genetically determined African ancestry and 12.2% for European ancestry. Updating the algorithm this way improved capture of T2DM-diagnosed individuals by 15.8%.

While updating the algorithm we used phenome and genome wide association study analyses to characterize cases and controls. We evaluated the value of various decision points that reflected clinical management. For example, by requiring control subjects to have a glucose or HbA1c measurement, which are not routinely measured in healthy patients without underlying comorbidities, we found that the early algorithm enriched the control population for patients with other metabolic dysfunction including a 140% increase in comorbid CKD, a 125% increase in comorbid hypertension, and a 217% increase in comorbid Hypertensive CKD compared to putative controls without a glucose test. Putative controls with a glucose test also had lower median HDL and higher median BMI and Triglyceride values. This inherent bias in lab order procedure could result in confounding bias that, impacts interpretability.

In conclusion, we demonstrate the necessity of updating phenotyping algorithms to improve capture. Additionally, we demonstrate how components of phenotyping algorithms can introduce false associations and diminish real ones.

References

Cohort Discovery with Unified Medical Language System Query Expansion

Robert J. Fawcett, MS\(^1\), Ahmad S. Halwani, MD\(^1\), Kyle R. Hansen, PhD\(^1\), Vikas V. Patil, MS\(^1\), Kelli M. Rasmussen, MD\(^1\), Daphne R. Friedman, MD\(^2\), Andrew R. Post, MD, PhD\(^1\)

\(^1\)Huntsman Cancer Institute, Salt Lake City, Utah; \(^2\)Duke University, Durham, North Carolina

Abstract

To simplify and enhance cancer cohort discovery, query expansion was employed by crawling the Unified Medical Language System’s (UMLS) concept tree to find International Classification of Disease for Oncology (ICDO) codes related to an initial code. Relevant ICDO codes were searched starting from a UMLS concept matched to a free-text query via natural language processing (NLP). We envision using this process to enhance a cancer cohort discovery tool that is under development.

Introduction

There is a strong desire to simplify self-service enterprise data warehouse searching by investigators for cohort discovery in oncology. NLP contributes to this simplification by allowing users to query data with language natural to their domain, but it has limitations. Simple NLP searches for concepts within the UMLS can produce associated medical codes (e.g. ICDO), but semantically related codes may also be useful to the searcher. In this work we used NLP to locate an initial set of ICDO codes, and then we traversed the UMLS concept tree to find related ICDO codes using narrower and broader relationships.

Methods

The UMLS dataset was installed into a MySQL database. QuickUMLS, an NLP Python application programming interface tool provided by the Georgetown University IR Lab, was installed for UMLS concept extraction from natural medical text. A python script accepted natural medical language, employed QuickUMLS to find associated ICDO codes from UMLS, and then queried the UMLS database using a recursive SQL call to traverse the concept tree \(x=2\) levels narrower and \(y=1\) levels broader, where \(x\) and \(y\) are adjustable parameters. The results were filtered to only those concepts associated with ICDO codes. The relevance of the ICDO codes returned were assessed by three domain experts who rated each of the 609 codes returned from a set of 22 initial search terms, with the following binary scale: 0=irrelevant, 1= relevant. For each search term, the rating results were averaged for a final percent relevancy score. Query performance was assessed by tracking the time taken, in milliseconds, to produce the final concept tree for each of the search terms tested.

Results

For the 22 search terms tested, the average percent relevancy score was 52% (max 100%, min 8%) with an average of 9.3 additional relevant codes returned per search term. There was good agreement among the reviewers, kappa = 0.66 (95% CI, 0.61 to 0.73). Query performance averaged 60.5ms from search term to ICDO tree.

Conclusion

The UMLS is composed of a variety of medical coding systems, controlled vocabularies and ontologies providing concept relationships not exhibited by using a single coding system in isolation. This research demonstrated that the UMLS concept tree can enhance cohort discovery by presenting the user with an expanded list of relevant medical codes. The process does produce noise, as only 52% of the results were relevant to the initial search term. More work needs to be done to discover causes of low performing search terms and identify associated UMLS features to increase the average percent relevancy score.

References

School District COVID-19 Response Policies Across Demographics

Ricky Flores, BS¹, Ellen Kerns, MPH², Russell J. McCulloh, MD², Aubree Honcoop, MPS², Martina A. Clarke, PhD¹,

¹School of Interdisciplinary Informatics, University of Nebraska-Omaha, Omaha, NE, USA; ²Department of Pediatrics, University of Nebraska Medical Center, Omaha, NE, USA

Introduction: With the COVID-19 pandemic, we have seen a rapid adoption of remote schooling practices nationwide which is inferior to in-school learning¹⁻². Children from communities who face disproportionate harm attributed to the pandemic may face even greater disparities due to remote learning; it is therefore important to understand how remote schooling practices align with populations and communities served by public schools nationally. This study’s objective is to evaluate the association between school pandemic response policies for districts in the United States and select publicly available student sociodemographic data.

Method: A cross sectional study design was used to evaluate the association of enrollment demographics and COVID-19 policies. Student demographic in addition to free and reduced lunch enrollment data was downloaded from each states’ respective Department of Education website. Thirty-three of the 50 states (66%) had publicly available demographic data along with the proportion of students from each school or district participating in a free and reduced lunch program or, in some cases, identified to be economically disadvantaged (defined as students and families who participate in federal assistance programs). Demographic data was composed of counts of students who fall into 1 of 8 categories for race/ethnicity: White, Black, Hispanic, Asian, Pacific Islander, American Indian/Alaskan Native, two or more races, or N/A. School District response to COVID-19 data was provided by MCH Strategic Data¹ and was joined to the demographic data via district name and state. MCH school data comes from responses gathered between July 2020 and January 2021. The proportion of students from each school who were identified as being part of a free and reduced lunch program or economically disadvantaged and the proportion identifying as being either Black or Hispanic were placed into 4 quartiles for ease of analysis. Chi-squared tests for association were performed for both sets of quartiles against the different COVID-19 responses.

Results:

Table 1: African American / Hispanic student quartiles and their association with COVID-19 response policies.

<table>
<thead>
<tr>
<th>COVID-19 Response Policy</th>
<th>Black / Hispanic students</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Q1: 0% - 11%</td>
</tr>
<tr>
<td>Teaching Method</td>
<td>Hybrid:</td>
</tr>
<tr>
<td></td>
<td>7902 (60%)</td>
</tr>
<tr>
<td></td>
<td>7420 (57%)</td>
</tr>
<tr>
<td></td>
<td>6724 (51%)</td>
</tr>
<tr>
<td></td>
<td>5800 (44%)</td>
</tr>
<tr>
<td></td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Staff Mask Policy</td>
<td>Required:</td>
</tr>
<tr>
<td></td>
<td>6206 (89%)</td>
</tr>
<tr>
<td></td>
<td>6454 (92%)</td>
</tr>
<tr>
<td></td>
<td>6609 (94%)</td>
</tr>
<tr>
<td></td>
<td>6868 (98%)</td>
</tr>
<tr>
<td></td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Sports Participation</td>
<td>Yes:</td>
</tr>
<tr>
<td></td>
<td>8460 (84%)</td>
</tr>
<tr>
<td></td>
<td>7672 (76%)</td>
</tr>
<tr>
<td></td>
<td>6983 (70%)</td>
</tr>
<tr>
<td></td>
<td>5855 (58%)</td>
</tr>
<tr>
<td></td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>

Table 2: Economically disadvantaged student quartiles and their association with COVID-19 response policies.

<table>
<thead>
<tr>
<th>COVID-19 Response Policy</th>
<th>Economically disadvantaged students</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Q1:0% - 32%</td>
</tr>
<tr>
<td>Teaching Method</td>
<td>Hybrid:</td>
</tr>
<tr>
<td></td>
<td>7194 (54%)</td>
</tr>
<tr>
<td></td>
<td>7252 (55%)</td>
</tr>
<tr>
<td></td>
<td>6810 (52%)</td>
</tr>
<tr>
<td></td>
<td>6667 (51%)</td>
</tr>
<tr>
<td></td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Staff Mask Policy</td>
<td>Required:</td>
</tr>
<tr>
<td></td>
<td>5184 (94%)</td>
</tr>
<tr>
<td></td>
<td>4984 (91%)</td>
</tr>
<tr>
<td></td>
<td>4960 (91%)</td>
</tr>
<tr>
<td></td>
<td>5135 (94%)</td>
</tr>
<tr>
<td></td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Sports Participation</td>
<td>Yes:</td>
</tr>
<tr>
<td></td>
<td>7222 (71%)</td>
</tr>
<tr>
<td></td>
<td>7753 (77%)</td>
</tr>
<tr>
<td></td>
<td>7349 (73%)</td>
</tr>
<tr>
<td></td>
<td>6726 (67%)</td>
</tr>
<tr>
<td></td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>

Conclusion: Schools with higher proportions of Black, Hispanic, and/or economically disadvantaged students more often offered only remote schooling options for their students. Schools serving more communities of color and economically disadvantaged students should be a target for efforts to increase safe in-person school attendance options.

COVID-19 concerns and interests differ with socioeconomic status: Twitter Analysis

Samah Fodeh,1,4 Aarthi Venkat,2 Yihua Su,1 Yadush Yadav,1 Lisa Puglisi3,4

1Health Informatics Program, Yale School of Public Health, 2Computational Biology and Bioinformatics Program, Yale University, 3Department of Internal Medicine, Yale School of Medicine, 4Department of Emergency Medicine, Yale School of Medicine

Abstract: The emergence of novel coronavirus disease (COVID-19) and its etiologic cause, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has prompted an international effort to limit its morbidity and mortality. We sought to understand how U.S. residents responded to COVID-19 as it emerged, and the extent to which socioeconomic status impacted materials and methods.

Materials and Methods:
We mined and reverse-geocoded 269,556 coronavirus-related tweets from January 23rd to March 25th, 2020. We then linked the tweets to a county-level area deprivation index (ADI) and areas with high initial disease counts (“hotspots”). We applied topic modeling1 to identify chief concerns. We first determined dominant topics and then computed the evolution of topics over time, and analyzed how topic proportions varied based on ADI and between hotspots and non-hotspots.

Results: We identified 45 topics, which shifted from early-outbreak-related content in January, to the presidential election and governmental response in February, to lifestyle changes in March. Highly resourced areas (low ADI) were concerned with stocks, social distancing, and national-level policies, while high ADI areas shared content with negative expression, prayers, and discussion of the CARES Act economic relief package. Topics had significance difference between both groups (p <.05) (see Figure 1). Within hotspots, these differences stand, with the addition of increased discussion regarding employment in high ADI versus low ADI hotspots (see Figure 2).

Conclusion: This work demonstrates a novel framework for assessing differential sentiments correlating to income, education, and housing disparities. This, with integration of COVID-19 hotspots, offers increased insight into differential topics and concerns inferred from the text in hotspots vs non-hotspots and in areas with low vs high ADI.

References:
Symptom Clustering of Patients with Suspected Acute Coronary Syndrome at Emergency Department Triage
Stephanie O. Frisch, PhD, RN\(^1\), Zeineb Bouzid, MS\(^1\), Jessica Zègre-Hemsey, PhD, RN\(^2\), Holli A. DeVon, PhD, RN, FAAN\(^3\), Harry Hochheiser, PhD\(^1\), and Ervin Sejdić, PhD\(^1\)
\(^1\)University of Pittsburgh, Pittsburgh, PA, USA; \(^2\)University of North Carolina at Chapel Hill, Chapel Hill, NC, USA; \(^3\)University of California Los Angeles, Los Angeles, CA, USA

**Introduction**: Accurate Emergency Department (ED) assessment of symptoms associated with acute coronary syndrome (ACS) or a heart attack is a driving force for timely critical treatments. Patients with a confirmed diagnosis of ACS report an average of 5 symptoms at ED triage. Agglomerative clustering is ideal for categorical data and has successfully been used on potential ACS symptoms. Understanding symptom clusters for these time sensitive events is key to developing evidence-based clinical decision-making tools to improve identification of patients with suspected ACS.

**Methods**: This was a retrospective observational cohort study of adult patients who present to the ED, at a regional health care system, with at least one symptom that was highly suspicious for ACS according to the American College of Cardiology guidelines. Independent reviewers extracted 24 symptoms from the electronic health record from the initial ED encounter. All chest pain and anginal-like symptoms were combined into one variable. We performed agglomerative hierarchical clustering with the following parameters: 1) clusters=2, 2) affinity=Euclidean, and 3) linkage=average. We computed the silhouette coefficient for each point and averaged it across all of the samples to determine the silhouette score to assess the clustering quality.

**Results**: Our sample included 1651 patients (age 64±14, 46% female, 10% black, 1% Hispanic). The top 5 most prevalent symptoms were: chest pain and anginal-like symptoms (64%), shortness of breath (54%), radiating pain (28%), feeling light headed (24%), and sweating/diaphoretic (18%). All other symptoms occurred in less than 18% of the sample. Twenty-four symptoms that were collected at initial ED encounter were analyzed using agglomerative hierarchical clustering with an average silhouette score of 0.346. Figure 1 represents the dendrogram with each color representing a cluster.

**Conclusion**: Agglomerative clustering had a fair ability to cluster symptoms of patients with suspected ACS at ED triage. Symptom clustering of patients with suspected ACS may be a useful technique for clustering symptoms or clustering individuals to improve risk stratification.

**References**
A Machine Learning Approach to Predicting Time Needed for Outpatient Surgery and Patient Recovery at a Freestanding Ambulatory Surgery Center

Authors: Rodney A. Gabriel, MD, MAS1,2 and Bhavya Harjai, BS3
1Department of Anesthesiology, University of California, San Diego, La Jolla, CA, USA
2Division of Biomedical Informatics, University of California, San Diego, La Jolla, CA, USA
3Department of Computer Science, Dayananda Sagar College of Engineering, Bangalore

Abstract: In order to more accurately predict whether outpatient surgeries at a freestanding ambulatory surgery center will be efficiently scheduled, we aimed to use machine learning to determine the probability a surgery will end and the patient will be discharged from the recovery room by the expected times.

Background: Freestanding ambulatory surgery centers (ASC) are facilities that perform outpatient surgeries and constitute a significant portion of total surgeries performed in the United States. Operating room efficiency at these centers is an important component of its financial model.1 Due to predetermined staffing of ancillary staff, nursing, and physicians, surgical procedures need to be completed by a set time (to avoid over-time pay for operating room staff) and patients need to be discharged from the recovery room at a set time (to avoid over-time pay for recovery room staff). For example, a typical time for surgical end time is by 5:00pm and for facility discharge time by 7:00pm. Thus, the sequence at which surgical procedures are scheduled in the operating room needs to be optimized so that all surgeries do not go pass 5:00pm and all patients are discharged before 7:00pm. Thus, in order to more accurately predict whether cases will be efficiently scheduled, we aimed to use machine learning to determine the probability a surgery will end and the patient will be discharged from the recovery room by the expected times.

Material and Methods: Our institutional review board (Human Research Protections Program) waived the need to consent for this study as it was not considered human subjects research. Data was collected for surgeries performed at our freestanding outpatient surgery center, including surgical procedure, surgeon estimation of case duration, surgeon, total operating room time estimation, surgeon experience, and patient data (American Society of Anesthesiologists classification score, age, body mass index, and sex). Furthermore, we collected data on actual operating room time required and post-anesthesia care unit (PACU) length of stay. We sought to determine whether a surgery was appropriate to book by a certain time (i.e. 3:00pm in this example). Given that all surgeries need to end by 5:00pm and patient needs to be discharged from PACU by 7:00pm, our binary output was Boolean value that was true if the actual operating room time was less than 120 mins AND the total perioperative time (operating room time plus PACU length of stay was less than 240 mins). We used the following machine learning algorithms (including all independent features into the models): 1) logistic regression, 2) random forest, 3) balanced-bagging, and 4) balanced-random forest. We utilized Synthetic Minority Oversampling (SMOTE) for oversampling the minority class. The metric used to validate the models was the F1-score, which is the harmonic mean of precision and recall. Stratified K-fold cross validation with 10 splits was set for 3 repetitions to get the F1 scores.

Results: There were 12,920 surgical procedures included in the analysis, comprising a mixture of orthopedic (35.6%), obstetrics/gynecology (14.3%), otolaryngology (13.2%), colorectal (8.9%), urology (6.9%), breast (6.5%), minimally invasive (5.6%), and other (8.9%) outpatient surgery. The mean actual total perioperative time (operating room duration plus PACU length of stay) was 182.9 minutes with standard deviation of 85.1 minutes. When predicting whether a surgical procedure may be appropriately booked by 3pm, the mean F1 scores for each machine learning algorithm was: 1) logistic regression = 0.698 and 0.772 (with SMOTE); 2) random forest = 0.811 and 0.943 (with SMOTE); 3) balanced random forest = 0.743 and 0.943 (with SMOTE); and 4) balanced bagging = 0.786 and 0.948 (with SMOTE).

Discussion: In our analysis, we found that when using machine learning approaches – including balanced random forest – in combination with SMOTE for oversampling minority classes, the F1 scores were improved. This type of analysis is important in operating room efficiency within a freestanding outpatient surgery center. In an effort to minimize overtime pay of resources/staff, it is vital that surgical procedures are booked at appropriate times. With these models, we can more accurately predict when surgeries can be booked at various times of the day. In our example, we chose 3:00pm as a case point; however, the model can be adjusted to predict booking appropriateness at various times of the day.

Telehealth Uptake and Continuing Usage During the COVID-19 Pandemic

Cheng Gao, PhD\textsuperscript{1}, Bradley A. Malin, PhD\textsuperscript{1,2}, You Chen, PhD\textsuperscript{1,2}

\textsuperscript{1}Vanderbilt University Medical Center, Nashville, TN; \textsuperscript{2}Vanderbilt University, Nashville, TN

Introduction

Telehealth has been adopted by providers to conduct safe and timely care for patients in response to the COVID-19 pandemic. Thus, the number of telehealth visits has exploded since March 2020. However, the change of telehealth usage in the COVID-19 pandemic and how different service departments use telehealth have never been explored. In this poster, the number of telehealth visits over a nearly one-year period is reported. In addition, We compare the differences in telehealth usage across several treatment facilities from the initial outbreak to the current time.

Methods

We identified telehealth, including audio and video, visits in the electronic health records (EHR) of patients at Vanderbilt University Medical Center (VUMC) between March 2020 to mid-February 2021. According to the trend of the number of telehealth visits, we partitioned the study period into two stages: 1) the beginning of the outbreak (03/01/2020 to 06/26/2020) and 2) continuing stage after the initial outbreak (06/27/2020 to 02/12/2020). We reported the average number of telehealth visits per day for these two stages in six treatment facilities, which had the highest average number of telehealth visits during the pandemic at our medical center. These facilities include Primary Care, Diabetes, Behavioral health, Rheumatology, Neurology, and Psychiatry. Also, the changes in terms of telehealth visits were calculated using \( \frac{\text{# of visits in stage 2} - \text{# of visits in stage 1}}{\text{# of visits in stage 1}} \).

Figure 1. The daily number of telehealth visits from 03/01/2020 to 02/12/2020

Results

Figure 1 reports on the total number number of telehealth visits per day. It can be seen that the volume increased exponentially during the first few weeks of the pandemic. The volume decreased from mid-April 2020 and became relatively steady in early July 2020. However, we observed a gradual increase of telehealth visits in December 2020, which coincides with the surge of COVID-19 cases in Tennessee during the same period.

The average number of telehealth visits for the outbreak and continuing stages for each of the facilities is shown in Table 1. Each of the facilities experienced a decrease in telehealth usage between the outbreak and continuing stage. However, there was a notable variation in such reductions. Four of the facilities, (Primary Care, Diabetes, Rheumatology, and Neurology) exhibit a dramatic drop of over 40%. By contrast, Behavioral health and Psychiatry facilities experienced only a minor reduction on the order of 10%.

Table 1. Changes in the average number of telehealth visits per day by different treatment facilities

<table>
<thead>
<tr>
<th>Treatment Facilities</th>
<th>Primary Care</th>
<th>Diabetes</th>
<th>Behavioral Health</th>
<th>Rheumatology</th>
<th>Neurology</th>
<th>Psychiatry</th>
</tr>
</thead>
<tbody>
<tr>
<td>03/01/2020 - 06/26/2020</td>
<td>173.6</td>
<td>79.6</td>
<td>57.4</td>
<td>69.6</td>
<td>65.3</td>
<td>33.7</td>
</tr>
<tr>
<td>06/27/2020 - 02/12/2021</td>
<td>55.7</td>
<td>45.9</td>
<td>52.4</td>
<td>37.6</td>
<td>28.7</td>
<td>29.8</td>
</tr>
<tr>
<td>Changes</td>
<td>-67.9%</td>
<td>-42.3%</td>
<td>-8.8%</td>
<td>-46.0%</td>
<td>-56.0%</td>
<td>-11.8%</td>
</tr>
</tbody>
</table>

Discussion

The mild reductions in telehealth usage between the outbreak to continuing stage indicate that telehealth may be more suitable for Behavioral health and Psychiatry. The reasons why a dramatic drop in telehealth usage transpired for certain treatment facilities (e.g., Primary Care) and not others are unclear. Possible causes include doctors’ concerns about care quality, reimbursement policy change, and access barriers for patients (e.g., lack of high-speed internet and hardware). Further investigation is suggested on these reasons.
Leveraging Assets-Oriented, Data-Driven Approaches in Global Health Informatics Curricula

Grace Gao, PhD, DNP, RN-BC1, Christie Martin, MN, MPH, RN, LHI-HP2, Michelle Wang, BS, BA3, Matthew Byrne, PhD, RN, CNE1, and Joyce Brettner, DNP, RN-BC1
1St Catherine University, St. Paul, Minnesota; 2School of Nursing, University of Minnesota, Minneapolis, Minnesota; 3University of Georgia, Athens, Georgia

Introduction
The COVID-19 pandemic has challenged and intensified social determinants of health (SDH) and social justice issues in global health communities. SDH indicators offer descriptors of different societies and are valuable as predictors of health outcomes relevant to population health worldwide, and as such, are practical and applicable in designing global health informatics curricula. To that end, these related topics can prepare health professionals to transfer seamlessly from education to workforce and from academic informatics education to applied health informatics practice. This poster provides an initial excursion into an innovative, assets-oriented, data-driven graduate global health informatics curriculum with a focus on SDH and social justice indicators in relation to countries’ income and geographical region using open-source databases and platforms including geographic information systems (GIS) maps and data.

Curricular Design Modeling
The curricula that model both global health and health informatics bridge disciplines in social science, health science, information science, cognitive science, and computer science. The proposed global health informatics curricular modeling is intended to bridge these different disciplines by leveraging health informatics as a means of serving the purpose of advancing the study, research, and practice of SDH and social justice in global health. The target audience for this global health informatics curricula is graduate interprofessional health students needing to meet global health informatics competency and skill development focused on SDH and social justice. By employing data-driven approaches that impact what is portrayed as global health, its past, present, and future, students can gain a better understanding of SDH factors and how these factors relate to one another and function in health systems within the context of various social and healthcare settings to discover health assets and detect health challenges facing global communities.

Assets-Oriented and Data-Driven Exemplar
A better understanding of SDH and health inequalities within and across populations can facilitate collaborations aimed at helping communities strategically identify and prioritize health issues, assets, and resources. The group learning activity in this exemplar presents an opportunity for interprofessional health students to collaborate in teams and use open-data sources to identify SDH and social justice factors that may shape health outcomes within the global setting. The learning activity is designed to employ an assets-oriented, and data-driven approach to select and analyze data to work through problems comparable to what interprofessional health students would encounter in real-life settings. Students can also use historical data and apply predictive data analytics to build predictive models. At the completion of this learning activity, students will understand factors that explain unequal distributions of health and disease in various populations worldwide and the role of interprofessional health teams play in addressing these issues.

Limitation and Challenges
Limitations can relate to issues with data integrity such as missing data, time constraints, and data currency, which are also common to real-time analyses of global health informatics data. It is critical to consider time constraints and design learning activities accordingly. In addition, selection bias in choosing data may also exist.

Conclusion
Open-source health data represents a valuable tool for teaching academic informatics since it is readily available and connects students to real-world data sources and settings. Leveraging open-source SDH and social justice data and GIS maps to discover potential health assets within countries by region can enhance students’ data analytic skills and competencies to address global health SDH concerns and promote health equity and social justice.
Application of a Pharmacy Alert Management Program to Reduce Nuisance Medication Clinical Decision Support Alerts

Sean S. Garcia, PharmD1, Vera L. Taylor, PharmBS1, Michael W. Mickan, MDiv, BSN1, Sandy Kyriakos, MBA, BSN1, Nnaemeka G. Okafor, MD, MS1
1Memorial Hermann Health System, Houston, Texas, United States

Introduction
Medication clinical decision support (mCDS) alerts assist clinicians with order entry and improve patient safety1. It has been reported, however, the majority of alerts are overridden by ordering providers, with override percentages of 96%1. If there is no institutional alert management program, alert fatigue occurs as a result of excessive, clinically insignificant alerts and diminishes the effectiveness of mCDS1. Several strategies of optimizing alerts to decrease alert burden on clinicians have been reported in the literature1. We report on our approach of using an informatics-led pharmacy alert management program to collaborate with pharmacist stakeholders and combat alert fatigue.

Methods
Dashboards available on Cerner’s Lights On Network (Kansas City, Missouri, United States) were utilized to identify mCDS alert types that were firing excessively at Memorial Hermann Health System, which operates 14 hospitals. The highest-volume alert types identified were targeted for optimization and alert suppression. Data was extracted from the Lights On Network to analyze the alert types further and determine the specific alerts that were firing to pharmacists. From August to October 2020, meetings led by an informatics pharmacist occurred monthly as needed to review mCDS alert data with pharmacy stakeholders from across the enterprise. Medication safety officers, pharmacists, pharmacy managers, and pharmacy directors were engaged for collaboration to gain consensus on which alerts were deemed nuisance alerts, which were defined as clinically insignificant alerts. Alerts identified as a nuisance were suppressed thereafter.

Results
The mCDS duplicate therapy (DT) alert and drug-drug interaction (DDI) alert types were observed firing the highest volume of alerts in our system, with peak DT alerts (42.4 alerts/100 orders; override rate (OR) 91.5%) in September 2020 and peak DDI alerts (29.9 alerts/100 orders; OR 96.8%) in October 2020. As pharmacist positions were receiving the most alerts in our system, the target population for alert suppression was pharmacists. DT nuisance alerts were suppressed in October 2020. From November 2020 to February 2021, 75,093 DT alerts per month on average were suppressed. The average monthly DT alert rate was 32.8 alerts/100 orders (22.6 % reduction) compared to the baseline DT alerts observed in September 2020. The average OR for the same time period is 89.65% and represents a 1.85% increase in the acceptance rate of DT alerts. Six of the highest-volume DDI alerts were suppressed in November 2020. From December 2020 to February 2021, 52,173 DDI alerts per month on average were suppressed. The average monthly DDI alert rate was 24.1 alerts/100 orders (19.4 % reduction) compared to the baseline DDI alerts observed in October 2020. The average OR for the same time period is 94.47% and represents a 0.33% increase in the acceptance rate of DDI alerts. Extrapolating a median alert dwell time of 7 seconds for pharmacists from prior studies, pharmacists across the enterprise are saved an estimated 146 hours (DT alerts) and 101 hours (DDI alerts) per month from suppressing nuisance alerts.2

Conclusion
A pharmacy alert management program, consisting of an informatics pharmacist routinely collaborating with pharmacist stakeholders, can allow institutions to target and refine mCDS duplicate therapy alerts, reduce nuisance alerts, modestly increase clinician acceptance rate, and recover hundreds of hours of pharmacist time per month.

References
Patient-Centered Care Practices in Secure Messaging: Qualitative Analysis of Interviews with VHA Primary Care Teams

Lynn A. Garvin, PhD,1,2 Emily Leonard, MPH,3 Linda Am, MPH,3 Timothy P. Hogan, PhD, 3,4 Stephanie L. Shimada, PhD2,3,5

1Center for Healthcare Organization and Implementation Research (CHOIR), VA Boston Healthcare System, Boston, MA; 2Boston University School of Public Health, Boston, MA; 3CHOIR, VA Bedford Healthcare System, Bedford, MA; 4Department of Population and Data Sciences, UT Southwestern Medical Center, Dallas, TX; 5University of Massachusetts Chan Medical School, Worcester, MA.

Objective: Understand secure messaging (SM) practices of primary care teams across the Veterans Health Administration (VHA). Identify challenges and practices that support patient-centered care (PCC). Methods: Thematic analysis applied to data from qualitative interviews with 20 VHA primary care team members. Results: Teams identified barriers and facilitators to SM use at VHA/facility-, team- and patient levels to build patient-provider relationships. Conclusions: Teams can incorporate SM into clinical practice to support PCC.

Introduction. Over the past decade, the Veterans Health Administration (VHA) has reorganized its primary care services into Patient Aligned Care Teams (PACTs) to deliver patient-centered care (PCC) and communication.1 Secure messaging (SM) through VHA’s patient portal My HealtheVet is one PCC tool. Studies have explored how SM supports PCC,3 yet little is known about how primary care SM practices enable or hinder PCC. This study’s purpose was to understand current VHA SM practices, and identify strong practices and improvement opportunities.

Methods. We conducted 20 semi-structured interviews with VHA PACT members (September 2020-March 2021), recruiting from VHA facilities nationwide among: primary care providers (PCPs), registered nurses (RNs), licensed practical nurses (LPNs) and other staff. Interviews were audio-recorded, transcribed, and double coded following principles of thematic analysis. Researchers identified factors impacting PACT SM use and strategies.

Results. (1) SM roles among PCPs and nurses. Team nurses triaged the messages received, delegating them to others for role-related action. PCPs principally responded to delegated SMs (e.g., authorizing prescription refills).
(2) SM work characteristics. Participants described prioritizing speed in responding to messages because patients report increased confidence in their care teams and peace of mind based on prompt replies. PCPs reported dedicating roughly 30 minutes per day to SM; RNs and LPNs spent approximately 30-60 minutes per day on SM.
(3) SM philosophy or mental checklist. Participants tried to strengthen quality of care through timely message responses, and through messages that engage patients in their care and not get “lost” to follow-up.
(4) SM protocols. Most participants start messages with a greeting that includes the patient’s name to build rapport. Participants said they matched the purpose and tone of the message in their response: brief and transactional or longer and personal. Most conclude their SM using the automatic signature block (with name and credential).
(5) SM in the patient-team relationship. Use of SM ensured a timely, reliable response that built patient trust in PACT clinicians and VHA. SM reduced unnecessary walk-ins and provided a record of the SM exchange.
(6) PCC in SM. Participants respected patient communication preferences (i.e., SM, phone call, or visit) while recognizing their responsibility to respond using the method appropriate to the patient’s health situation.
(7) Types of SMs that get special handling. If patient messages presented medical, mental health or emotional complexity or urgency, participants elevated their response (via SM, phone call, appointment scheduled).
(8) Improvements to SM at organizational, team and individual levels. Improvement recommendations ranged from IT communication on SM system downtime, to interoperability and user interface, to patient training.

Discussion. VHA’s focus on PCC practices in SM is reflected in primary care teams’ efforts to improve patient access to their health information and patient experience with their care. SM is an integral component of PCC.

References
A Comparative Time-to-Event Analysis Across Health Systems

Mohamed Ghalwash, PhD1, Prithwish Chakraborty, PhD1, Akira Koseki, PhD2, Hiroki Yanagisawa, PhD2, Toshiya Iwamori, MS2, Reitaro Tokumasu, PhD2, Masaki Makino, MD, Ph.D3, Ryosuke Yanagiya, AA4, Atsushi Suzuki, MD, Ph.D3, Michiharu Kudo, PhD2, Daby Sow, PhD1

1IBM Research, Center for Computational Health, NY, USA. 2IBM Research, Tokyo, Japan. 3Fujita Health University, Department of Endocrinology, Diabetes and Metabolism, Aichi, Japan. 4Fujita Health University, Division of Medical Information Systems, Aichi, Japan.

While the wide spread adoption of EHR allows access to clinical data for AI modeling, constructed cohorts at institutional levels tend to be small due to inevitable data sparsity issues, which poses a challenge to build interoperable models performing well across health systems. This work addresses such interoperability challenges, focusing on the development of time-to-event models to estimate Type 2 Diabetes Mellitus (T2D) complication times. To investigate the interoperability of such models across health systems, we applied state-of-the-art survival models on cohorts from two different health systems: US and Japan. The US cohort is extracted from a large claims dataset spanning 5 million enrollees. The Japanese cohort is extracted from Fujita Health University Hospital (FHUH). In both cohorts, we use the first T2D diagnosis as the index event and model the onset of Neuropathy. The US (FHUH) cohort covers 6981 (123) enrollees, of whom 232 (41) developed Neuropathy, respectively.

![Figure 1: Interoperability across health systems](image)

Deep survival (DeepSurv) and variational autoencoder Cox (VAECox) are two of the state-of-the-art methods and we used these two to study model interoperability. The concordance index (CI) results of applying these methods on our cohort are shown in Figure 1. The models trained on FHUH show less generalization when evaluated on the US cohort. For example, the CI of the DeepSurv model when trained and evaluated on FHUH is 0.907. However, when evaluating the pre-trained FHUH model on the US cohort the CI dropped to 0.543. This might indicate that the model trained on FHUH can not be generalized to other cohorts. On the other hand, the CI of DeepSurv when trained and evaluated on the US cohort is 0.605. Also, when evaluating the pre-trained US DeepSurv model on FHUH the CI does not drop down. Similar finding using the VAECox were obtained. Figure 1 shows the important features contributing to the estimation of the pre-trained US DeepSurv model using SHAP. Each dot represents one patient; higher values for the features are colored in red and lower values are colored in blue. The results show that having other nervous system disorder increases the risk of Neuropathy. The model shows that male (blue dots for SEX) are more prone to Neuropathy than female. As part of the ongoing analysis, we plan to investigate why patients with Disorder of lipid metabolism have lower risk of Neuropathy. We are also studying various transfer learning techniques to address such model interoperability challenges.

References

Utilizing Advanced Visualization Technology to Study Home Care Challenges

Denise Goldsmith MPH, MS, RN, FAAN, Sara Flash BS, Jim Holdnack PhD
Patricia Flatley Brennan RN, PhD, FAAN
National Institute of Nursing Research, Washington, DC

Introduction
Digital technologies have dramatically impacted patient care with the advent of electronic medical records, wearable medical devices, and big data analytics. The Advanced Visualization Branch (AVB) of the National Institute of Nursing Research (NINR) evaluates the potential for advanced digital visualization technologies to understand the personal and environmental challenges of homecare and self-management. These technologies offer the potential to study common self-care and home-care activities in a controlled environment. As robust data collection and evaluation is a critical element of digital research platforms, we are beginning to develop informatics guidance to characterize advanced visualization for research and intervention. We use immersive virtual reality (IVR) to study how people carry out instrumental activities of self-care management, the impact of cognitive challenges and distractors on performance, and the way observations may complement or supplant self-report data.

Method
The AVB develops interactive IVR simulations that present information to patients with a variety of complex health conditions. These simulations enable observation and measurement of problem-solving behaviors that may impact health outcomes within home-based care environments. This research evaluates digital technologies as a tool to aid in the assessment and management of factors that impact the transition from acute care settings to outpatient environments. Virtual reality environments enable us to build standardized natural settings with experimental controls that would be nearly impossible to perform in real life. We can unobtrusively observe real-time decision making so that the measurement does not perturb the natural behavior of the participant and are able to modify behavior by giving immediate behavioral feedback during intervention trials. We utilize informatics principles of naming conventions, structured information labels and written descriptions of objects, including metadata that supports mapping of virtual object’s physical parameters. Large volumes of data are captured and exported as flat files for later statistical analysis.

Results
Our first simulation investigates how individuals behave in a virtual grocery store. We are inspired by the challenge of assessing patients’ ability to follow sodium restricted diets. In the simulation, the user navigates through a typical grocery store, reads nutrition labels with sodium information, selects food as if for purchase, and places food into a basket. We measure behaviors associated with and indicative of cognitive process, such as frequency of label referencing, product comparison, search strategy and completion of selected tasks common to a shopping trip. At checkout, the selected parameters, such as dietary sodium content, are calculated and users receive visual feedback.

Discussion
We utilize a consistent approach to presenting sodium content, use formal terms to represent nutritional values and common units of measure. Some of these are used to systematically plan store layout while others are presented to the participant. We apply principles of formalization, standards, and nomenclature to characterize this novel environment. This exploratory work leads us to interesting questions: can we capture and realistically represent environmental conditions that disrupt successful self-care behavior, and can we unobtrusively measure individual differences that contribute to successful outcomes?

References
1. Werner NE, Jolliff AF, Casper G, Martell T, Ponto K. Home is where the head is: a distributed cognition account of personal health information management in the home among those with chronic illness. Ergonomics, 2018; 61:8, 1065-1078.
Self-Reported Symptoms for COVID-19 Public Health Surveillance: A Window to Social Determinants of Health

Hope G. Gray, MTS, RBC, Mohanraj Thirumalai, PhD, Wayne H. Liang, MD, MS, James J. Cimino, MD, FACMI, FACP, FAMIA, Sue S. Feldman, RN, MEd, PhD
University of Alabama at Birmingham

Introduction
In response to the COVID-19 pandemic, we launched HelpBeatCOVID19.org, a novel self-reporting symptom surveillance system based at the University of Alabama at Birmingham and focused on Alabama and the Southeast region of the United States\(^1\). Helpbeatcovid19.org captures social determinants of health (SDOH) data and is focused on a region of the USA that is particularly vulnerable to the potential sequelae of COVID-19. This study concentrates on Jefferson County, Alabama, which is 43.5% Black, has a 2018 CDC Social Vulnerability Score of 0.6621 and a high score of 0.899 for the Race/Ethnicity/Language Theme\(^2,3\). This presentation will report on research in progress to understand the utility of self-reported data with communicable disease outbreaks.

Methods
Individuals voluntarily completed an online questionnaire at HelpBeatCOVID19.org which captured SDOH data and other disease surveillance variables. The data are stored on HIPAA-compliant servers. De-identified self-reported data were culled from the HelpBeatCOVID19 database, cleaned, sorted, and analyzed by zip code. Using STATA/SE 16.1, we used regression analysis to understand associations based on zip codes, especially where there are health disparities in historically African American neighborhoods in Jefferson County. Comparisons were also made with Alabama Department of Public Health COVID-19 case reports by zip code.

Results
To date there are 102,308 people who have reported their symptoms in HelpBeatCOVID19. Of those, 77,903 are from Alabama. Midfield, AL, a predominantly African-American neighborhood (81.1\(^4\)), has 74.1% of people reporting underlying conditions where the median household income is $38,750\(^4\). By comparison, Vestavia Hills, AL, a more affluent neighborhood with 88.8% White population and median household income being $109,485, had more people participating in HelpBeatCOVID19 (3,920), yet a smaller percentage (15.2%) with underlying health conditions.

Discussion
Our intermediate analysis of the data reveals that in Jefferson County, AL, a greater number of people in more affluent communities participated in the study. Whereas state-wide, a greater percent of individuals indicated that they had zero symptoms. Limitations of this study include the unclear reliability of self-reported data, and incorrectly entered data (such as phone numbers entered in a zip code field). Although many people have smart phones, we are uncertain how the lack of accessibility to technology and reliable internet connectivity could impact the use of web based surveillance systems.

Conclusion
We believe that future studies can include additional questions to help capture SDOH and their impact on health disparities. This study begs the question: It is possible that people under reported their underlying conditions while using HelpBeatCOVID19.org and would benefit from further investigation.

Acknowledgements
This study was supported by the National Center for Advancing Translational Sciences of the National Institutes of Health under award number TL1TR003106. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References
Design, Development, and Usability of a Hypertension Medication Self-Management Conversational Agent (Medicagent)

Ashley C. Griffin, PhD, MPSH\textsuperscript{1,2}, Stacy Bailey, PhD, MPH\textsuperscript{3},
Saif Khairat, PhD, MPH, FAMIA\textsuperscript{1,2}, Yue Wang, PhD\textsuperscript{1,2},
Jaime Arguello, PhD, MS\textsuperscript{1}, Arlene E. Chung, MD, MHA, MMCi, FAMIA\textsuperscript{1,2}
\textsuperscript{1}University of North Carolina at Chapel Hill, Chapel Hill, NC, USA; \textsuperscript{2}Carolina Health Informatics Program, Chapel Hill, NC, USA; \textsuperscript{3}Northwestern University, Chicago, IL, USA

Introduction

Hypertension is the most common chronic disease in the U.S. and leading risk factor for heart disease\textsuperscript{1}. While approximately 75\% of adults with hypertension are taking medications, only half have adequately controlled blood pressures\textsuperscript{1}. Conversational agents, which are systems that converse with people, have demonstrated early potential to deliver interactive self-management interventions. Very little research has focused on conversational agents for hypertension. We leveraged a user-centered design process to design and develop, and evaluate the usability of a conversational agent focused on hypertension medication self-management called “Medicagent.”

Methods

Features of Medicagent were informed by semi-structured interviews from our prior study of patients’ needs and preferences on topics such as medications, refills, communication with the care team, and resources. Based on these needs, a functional, text-based prototype of Medicagent was iteratively designed and developed using Google Cloud’s Dialogflow natural language understanding engine. Dialogflow utilizes machine learning to match natural language utterances to intents (i.e., goal of the user’s query). The prototype underwent extensive pilot testing from health professionals and informatics researchers. We iteratively added utterances identified from chat logs to the training phrases (e.g., “view medication list” or “add my blood pressure”). For usability testing, purposive sampling was used to select ten adults with hypertension taking at least one prescribed medication based on age, race, gender, education, and number of medications. Usability testing sessions were conducted on a web-based version of Medicagent on participants’ personal computers during virtual video-recorded sessions. The session included: 1) background questionnaire, 2) ten hypothetical tasks within Medicagent using concurrent think aloud methods, 3) System Usability Scale (SUS) questionnaire, and 4) a brief semi-structured interview. Effectiveness, efficiency, and satisfaction metrics in the International Organization for Standardization 9241-11 standards were used to assess usability of the prototype.

Results

Ten participants completed the usability testing session. The average age was 60 years, 50\% were female, 50\% were Black, and 50\% had at least a college education. Participants were taking on average four medications. Participants spent a mean of 18 minutes (SD=10) interacting with Medicagent during the testing session, and nearly all tasks (98\%, 98/100) were completed. Two participants made errors inputting the correct medication and confirming that a new medication was added to the medication list. 8.6\% (11/128) of utterances were not successfully mapped to an intent. These errors resulted from unrecognized spelling or formatting of dates, times, and blood pressure values. Medicagent achieved a mean SUS score of 78.8/100, which demonstrates acceptable usability. Several participants had difficulties navigating the conversational interface without menu and back buttons, felt additional information would be useful for redirection when utterances were not recognized, and reported the need to provide a health professional persona.

Conclusion

In the emergent field of conversational agents for self-management, this study is one of the first to describe the design, development, and usability of a conversational agent focused on hypertension medication self-management. We identified areas for refinement, which could be used to inform the subsequent phase in the user-centered design process towards evaluating potential efficacy. Additional usability research for conversational agents should investigate the appropriateness of responses and unintended consequences of errors within the context of self-management.

References

Digital Methodology for Mobile Clinical Decision Support Development

Amy Grimsley, MSN, RN, CCRN-K1, Katrina Boles, MS1, Chelsea Howland, MSN1, Chuka Emezue, PhD, MPH, MPA1, Malaika R. Gallimore, RN, MPH1, LeeAnne B. Sherwin, PhD, MS, FNP-BC1, Jo-Ana D. Chase, PhD1, Blaine Reeder, PhD1, Allison B. Anbari PhD, RN1

1University of Missouri, Columbia, MO, USA

Introduction
Research at academic institutions faced a number of barriers during the global pandemic when in-person interactions were limited to prevent the spread of the COVID-19 virus. Research activities to develop a mobile clinical decision support app in the Precision Smart Technologies and Application for Rapid Translation (Precision START) laboratory at the University of Missouri quickly pivoted from a traditional face-to-face methodology to a fully digital methodology and maintained the project momentum. Our Mobile Application Information System for Integrated Evidence (MAISIE) project to bring evidence-based apps to nursing homes (NH) launched right before stay-at-home orders in March 2020. MAISIE extends prior work that developed a mobile clinical decision support app for evidence-based urinary tract infection (UTI) symptom concordance and anti-microbial stewardship in NH1.

Methods
The MAISIE team is homed in the Precision START laboratory. Lab membership totals nine members (five graduate students and four faculty) in five states. Our digital methodology implements methods translated from previous experience gained in applied technology research in NH and other contexts. Specifically, we pivoted from in-person meetings to video meetings for all lab activities including weekly team, stakeholder, and software developer meetings. App feature sets were accessed via the web when posted by the developer. Usability and function testing were conducted remotely. Test results were aggregated and discussed online.

Results
The MAISIE team has met exclusively online since March 2020 with limited interruptions to project progress. The creation of new case scenarios and interface design recommendations extended previous work. The MAISIE app was developed with interfaces that follow a head-to-toe assessment workflow typical of nursing practice (Figure 1). In addition to implementing a UTI symptom checklist based on prior work1, we added a COVID-19 symptom checklist feature based on guidance from the Centers for Disease Control and the World Health Organization. The app is ready for testing with external stakeholders and nurses in NH settings who will download and install the app from a development website and conduct usability tests via synchronous remote testing.

Conclusion
A completely digital methodology enables successful design, development, and testing of mobile clinical decision support apps. Advantages of this approach are minimization of project delays, inclusion of remote research team members, and elimination of infection risk for team members, software developers, external advisors, NH staff and residents. This digital approach can be applied to any health-related mobile app regardless of setting or target user. Next steps are to implement features for the five most common conditions that lead to preventable hospitalization for NH residents. Full description of methodology and current evaluation results will be presented at the conference poster session. Evaluation metrics will include participant usability and technical testing.

References
Impact of Comorbidity Profiles on Pain Trajectories in Breast Cancer Patients by Using Electronic Health Record Data

Jia-Wen Guo, PhD1, Katherine A. Sward, PhD1, Ann Lyons, PhD1, Susan L. Beck, PhD1,
Gary W. Donaldson, PhD1, Wendy W. Chapman, PhD2, Lewis J. Frey, PhD3
1University of Utah, Salt Lake City, Utah, USA; 2University of Melbourne, Australia;
3Medical University of South Carolina, Charleston, South Carolina, USA

Introduction

Approximately 20% to 30% of the breast cancer patients reported different levels of pain,1 despite a variety of treatments. According to the Integrated Model of Multimorbidity and Symptom Science,2 comorbidities, or the coexisting conditions in addition to a primary disease of interest, are part of the contributing or risk factors to patient outcomes. The impact of comorbidities on pain in breast cancer patients has been little studied. Therefore, the purpose of this study was to examine the association between pain trajectories of breast cancer and the comorbidity profiles.

Methods

This is a secondary data analysis from a retrospective, observational cohort study using electronic health record (EHR) data from a single medical center. In the primary study, four types of comorbidity profiles, based on the 17-items Charlson comorbidity index, were identified from 4412 adult inpatients (age>18 years) of any kind of cancer diagnosed from 12/2010 to 05/2019. In this study, only breast cancer (N= 1740) was included from the primary study dataset. The comorbidity profiles of the sample were: few comorbidities with nonmetastatic cancer (68.8%), few comorbidities with metastatic cancer (24.3%), multiple comorbidities with nonmetastatic cancer (4.4%), and multiple comorbidities with metastatic cancer (2.5%). To identify distinct longitudinal patterns of pain trajectory, the patients’ first two-year pain score data (0=“no pain”, 10=“worst pain possible”) since the first admission was used by group-based trajectory modeling (GBTM). Then, Multinomial logistic regression modeled the comorbidity profiles as predictors of pain trajectory patterns, including age and cancer stage as covariates.

Results

The mean age of the sample was 57.8 years (SD=12.7, range=20.9-96.7). The majority of the sample were female (99.2%), non-Hispanic/Latino (94.2%), White (90.3%), and a cancer stage of <=2 (79.7%). Four pain trajectory patterns were identified by GBTM: stable/consistent very mild pain (79.0%), increasing from mild to moderate pain (8.4%), decreasing from mild to very mild pain (7.8%), and consistent moderate pain (4.7%). When comparing patients’ demographic and clinical characteristics among different pain trajectory patterns, patients in consistent moderate pain were more likely to be at the late cancer stage (p<.001). The multinomial logistic regression analysis compared other trajectories to the stable/consistent very mild pain pattern; the comorbidity profiles were the significant predictors in all the pain trajectory patterns. Moreover, the cancer stage was another predictor in the trajectory decreasing from mild to very mild pain (OR:1.28, 95% CI: 1.06–1.56, p=.012) and the consistent moderate pain trajectory (OR:1.44, 95% CI: 1.13–1.84, p=.003). Age was not a significant predictor for any of the trajectory patterns.

Conclusions

Identifying risk factor profiles shared by specific subgroups of patients provides the first step to assist clinicians in caring for patients with complex cancer pain; however, this study was constrained by the limits of the EHR data included in the primary study. Clinically, patients in increasing from mild to moderate pain and consistent moderate pain trajectory could benefit from provider awareness of the trajectory for better pain management. Further studies by including other variables and taking consideration of different comorbidity profiles will assist in developing tailored interventions for patients with a specific pain trajectory.

References


Acknowledgment

This project is funded by a K01 award from the NIH National Institute of Nursing Research (K01NR016948, PI: Guo).
YouTube Video Analytics for COVID-19 Literacy
Yawan Guo, MS¹, Xiao Liu, PhD², Anjana Susarla, PhD³, Rema Padman, PhD¹
¹Carnegie Mellon University, Pittsburgh, PA; ²Arizona State University, Tempe, AZ; ³Michigan State University, Lansing, MI

Introduction
The rapid evolution of COVID-19 into a pandemic has called for improved health literacy to inform and educate the public.¹ YouTube, the largest video-sharing social media platform, hosts several million health-related videos about the pathogenesis, diagnosis, treatment, and prevention of various medical conditions and provides publicly available, user-generated health information in a rich visual format, which may be easier to understand and comply with for patients at all levels of health literacy. However, YouTube video contents may also misinform and contradict reference guidelines and there is lack of a clear and consistent mechanism to retrieve high-quality videos for health education.² This preliminary study develops an automated solution to filter and evaluate YouTube videos on COVID-19 topics from health literacy perspectives that consumers can access on a rapidly evolving and dangerous pandemic.

Methods
Our data collection process generates keywords about COVID-19 and retrieves videos matching these keywords, along with their metadata, using YouTube Data API. Keywords are collected from WHO’s and CDC’s FAQ pages, and online forums where consumers pose relevant search terms for information they seek. We extract features from the collected videos to assess their medical information content and understandability. Videos are annotated for understandability using AHRQ’s Patient Education Materials Assessment Tool (PEMAT)³ for audio-visual materials by graduate students with medical and non-medical backgrounds to represent clinical and consumer perspectives. Google Cloud Video Intelligence API functions such as object detection, video transcription, scene detection and optical character recognition are applied to extract important video features to automate PEMAT constructs. The Unified Medical Language System acts as a reference for medical information annotation and used to train a Bidirectional Long Short-Term Memory Recurrent Neural Network to extract medical terms from video descriptions and assess medical content. Finally, medical content and understandability features are used to train and evaluate several machine learning classifiers to recommend videos for patient education on COVID-19 related topics.

Results
We fetch the top 25 videos, contributed by both individual users and reputable healthcare organizations, for 40 search terms and store their ranking and metadata for further analysis. Limiting to videos of short duration (1-6 minutes) for easy viewing resulted in 304 unique videos for annotation. 33 features are extracted from the videos for building different classifiers for video recommendation classification. The models are trained on 228 randomly selected videos and evaluated on 76 hold out videos using four widely used machine learning methods. Results in Table 1 show the Random Forest classifier exhibiting the best performance. In addition to the number of unique medical terms extracted from the video narratives by the Bi-LSTM deep learning model, indicating the importance of medical content, video engagement measures such as duration, favorites and comments are also significant predictors for the recommended videos. 24 videos are in the final recommended set.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Forest</td>
<td>0.5667</td>
<td>0.68</td>
<td>0.6182</td>
<td>0.7237</td>
</tr>
<tr>
<td>SVM</td>
<td>0.5882</td>
<td>0.4</td>
<td>0.4762</td>
<td>0.7105</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>0.4118</td>
<td>0.412</td>
<td>0.412</td>
<td>0.6721</td>
</tr>
<tr>
<td>AdaBoost</td>
<td>0.5294</td>
<td>0.72</td>
<td>0.6102</td>
<td>0.6515</td>
</tr>
</tbody>
</table>

Table 1. Video Recommendation Classification Results

Discussion
In this study, we synthesize information retrieval, machine learning and statistical methods to identify YouTube videos on COVID-19 related topics that encode understandable medical content. Our findings have implications for consumers, healthcare and public health professionals, patient educators, researchers and policymakers. Next steps include extracting additional features to help delineate specific infectious disease related topics, improving classification performance, adding criteria such as content accuracy and evaluation by clinical experts and consumers.

References
Parkinson and movement disorders ontology for clinically-oriented and clinicians-driven data mining of multi-center cohorts in Parkinson's disease.

Deepak K. Gupta, MD1, Massimo Marano, MD2, Raj Aurora1, James Boyd, MD1-Satya S. Sahoo, PhD3

1Department of Neurological Sciences, Larner College of Medicine University of Vermont, 2Unit of Neurology, Neurophysiology and Neurobiology, Department of Medicine, University Campus Bio-Medico of Rome, Rome, Italy
2Department of Population and Quantitative Health Sciences, School of Medicine, Case Western Reserve University, Cleveland, OH

Movement disorders is a subspecialty of neurology, and includes several disabling, common and rare neurological disorders, including Parkinson's disease (PD), essential tremor, dystonia, Huntington's disease etc. PD is the most well-known of all movement disorders(1), and related to PD, there is a group of rare but rapidly progressive and even more disabling disorders, collectively termed as atypical parkinsonian disorders (APD). Although there are distinct clinical and pathological diagnostic criterion for PD and APD, patients present in the clinic with overlapping clinical features, which evolve with a great deal of variability and complexity over time. This leads to high level of uncertainty in the prediction of diagnosis and progression for an individual patient on clinical grounds. These challenges faced by patients, their families and clinician are further exacerbated by the lack of sensitive or specific diagnostics tests and biomarkers. Therefore, there is a clear need for a clinical decision support tool that can integrate and organize relevant data for informed decision making. A key critical gap in this respect is the lack of a comprehensive domain ontology for movement disorders, which can enable formal knowledge model for data input, semantic data integration, and support analytical queries for predictive modeling in parkinsonian disorders.

Although an ontology for Parkinson’s Disease (PDON)(2) has been developed, the focus of PDON is on basic science research that significantly limits it use in clinical as well as clinical research data query and retrieval from PD databases and biorepositories. In this abstract, we introduce the Parkinson and Movement Disorders Ontology (PMDO) that has been designed to: (i) Improve access to both research and clinical trial data being made available by open-access PD databases, such as Accelerating Medicine Partnership in Parkinson Disease (AMP PD) and Parkinson Progression Marker Initiative (PPMI), and clinical trials databases maintained by the Parkinsonian Study Group (PSG); (ii) Enable interoperability and mappings between a variety of instrument/scales that are used in the evaluation of patients in movement disorders clinical centers; and (iii) Facilitate multi-institution research studies through the development of interoperable data management tools with adequate statistical power.

The development of PMDO was based on three criteria: (i) literature survey on the phenotypic and genotypic characteristics of parkinsonian disorder; (ii) diagnostic criteria used for patient evaluation; (iii) metadata as well as data attributes of research cohorts, including longitudinal study methods and cohort inclusion exclusion criteria. The PMDO was developed using an iterative collaborative process with the use of clinical workflow used by movement disorders specialists as reference protocol for prediction of diagnosis, progression, and prognosis, as we all as ontology engineering best practices that included use of structural metrics to evaluate the information content of the ontology and identify modeling errors and improve the structure of class hierarchy. The PMDO class hierarchy uses the Basic Formal Ontology (BFO) as a reference upper-level ontology. The current version of PMDO consists of terms representing the phenotypic traits of PD and APD, the components of diagnostic criteria used during patient evaluation, and data elements used in the PPMI and AMP PD cohort studies, which are modeled as subclasses of the three BFO classes of DependentContinuant, IndependentContinuant, and Occurrent.

We are currently expanding PMDO to include all relevant clinical and research terms of all movement disorders, with a particular focus on modelling the complexity of the underlying pathophysiological characteristics that lead to disparate phenotypes in patients, with an emphasis on modeling genotypic data with explicit correlation with symptoms exhibited by patients. In addition, there is a clear need to model the neuropathology findings in terms of fluid and imaging biomarkers. As for application of PMDO, we are integrating PMDO with the Insight system(3) and the PPMI database for creating Insight-PD, an Ontology-driven data integration and patient cohort query platform for retrospective datamining of such databases. Furthermore, PMDO will form the foundation of Ontology-based, Real-time, Machine learning Informatics System for Parkinson’s Disease (ORMIS-PD) clinical informatics system for predicting PD subtype and prognosis at point-of-care in clinic. The PMDO is currently available at BioPortal at following link: https://bioportal.bioontology.org/ontologies/PMDO.

References
A Mixed-Methods Evaluation of Telehealth Adoption in Frontier Critical Access Hospitals

Saira N. Haque, PhD, MHSA, FAMIA, Alison Banger, MPH, Sydney DeStefano, BS, Regina Rutledge, PhD, Melissa Romaire, PhD
RTI International, Research Triangle Park, NC

Introduction

Frontier areas have low population density at great geographic distance from population centers and services. As a result, the Critical Access Hospitals (CAHs) that serve these areas might not be able to provide the services that are needed in the community. One way to ameliorate these challenges is through telehealth. Policy and reimbursement considerations are among the most significant barriers to greater telehealth uptake. Thus, the Center for Medicare and Medicaid Services (CMS) embarked on the Frontier Community Health Integration Project Demonstration (FCHIP) from August 2016 through July 2019 to learn how changing payment for telehealth to cost-based reimbursement impacted telehealth uptake and use. In particular, we were interested in impacts on hospital administration, hospital finances, patient/consumer access and spillover effects in the community.

Methods

We conducted a mixed-methods evaluation of the 8 hospitals in Montana, Nevada, and North Dakota that participated in the telehealth component of the FCHIP demonstration. The study was reviewed by the RTI International Institutional Review Board. Quantitative methods involved comparing frequencies and types of telehealth claims of Medicare beneficiaries who were part of the FCHIP demonstration versus a comparison group of beneficiaries in non-FCHIP CAHs in Montana, Nevada, and North Dakota. Primary analyses were hospital-based frequencies and secondary were beneficiary-based differences. Qualitative methods were key informant interviews and document review. We developed semi-structured interview guides that were tailored to common roles across FCHIP CAHs (e.g., provider, telehealth coordinator, administrator). We conducted annual in-person or telephonic interviews through the duration of the demonstration, which were recorded and transcribed. In addition, we reviewed program documentation that the FCHIP CAHs provided. Transcripts and documents were coded using NVIVO to support thematic analysis for all four domains of interest (hospital administration, hospital finances, access and spillover effects).

Results

We found varied results across the four domains of interest. For hospital administration, the FCHIP CAHs used technical support and assistance provided to demonstration participants to inform telehealth implementation, marketing and outreach. Implementation support included changes to clinical and administrative workflows and training and identifying clinical and administrative champions to facilitate workflow changes and coordination between the sites. Claims volumes were low, which resulted in nominal impacts on hospital finances. By the end of the demonstration, 6 out of the 8 participating sites billed Medicare for telehealth services, although all reported providing telehealth services. Some CAH administrators noted positive impacts on hospital finances due to ancillary services being conducted at the CAH at the direction of distant sites. The 6 FCHIP CAHs that billed Medicare had 289 telehealth encounters for 150 unique beneficiaries. The specialists that were accessed the most were cardiology, nephrology and behavioral health. CAHs reported that their patient population appreciated not having to travel great distances to receive care. Spillover effects such as impacts on the surrounding regional health delivery system, providers of community-based services and payers, varied across CAHs. The variation was based on factors such as relationships with neighboring hospitals.

Conclusion

The frontier setting is characterized by low population, and thus the volumes of telehealth services provided in the CAHs and comparison sites are low. Overall, CAHs reported that patient satisfaction was high and expressed the desire for more virtual services. Telehealth service selection was informed by community needs, but formal community needs assessments were not completed. Future work could include standardized, formal community needs assessments and assistance finding distant providers to meet those needs. Implementation support services helped CAHs integrate telehealth into clinical and operational workflows. However, billed telehealth services were fewer than reported services. This indicates a need for assistance around billing.

References

Enhancing Clinical Relevance of Health Behavior Insights via Semantics

Jonathan Harris¹, Deborah L. McGuinness, Ph.D.¹, Marco Monti, Ph.D.², Oshani Seneviratne, Ph.D.¹, Mohammed J. Zaki, Ph.D.¹, Ching-Hua Chen, Ph.D.³
¹Rensselaer Polytechnic Institute, Troy, NY; ²IBM, Italy; ³Center for Computational Health, IBM Research, Yorktown Heights, NY

Introduction
Personal health applications are designed to help users adopt healthier lifestyles. In general, however, it has been challenging to bring consumer health applications into clinical settings. The insights provided to users for tracking and monitoring their progress are not always immediately related to clinically meaningful concepts. To enable richer and more meaningful summaries, we extended an existing Time-Series Summarization (TSS) framework to dynamically generate semantic representations of frequent and/or anomalous behavioral patterns discovered from personal health data that is tracked over time. These behavioral patterns are used to populate a personal health knowledge graph (PHKG). To improve clinical relevance, we implemented a semantic reasoner that reasons over the PHKG as well as semantic rules based on health guidelines from the American Diabetes Association (ADA).

Method
Based on interviews with 21 adults with type 2 diabetes, we defined a Personal Health Ontology (PHO) to capture behavioral concepts (e.g., dietary patterns and preferences) and relationships to be instantiated in the PHKG. The TSS generates Resource Description Framework (RDF) triples to populate the PHKG with summaries of the user’s dietary behaviors. The PHKG has been instantiated as a queryable Blazegraph™ database. Additionally, several ADA guidelines related to diet and activity were modeled in a computable form using the Web Ontology Language (OWL). A semantic reasoner performed logical inference over the PHKG, the guideline rules, and an existing Food KG (which contains knowledge of hundreds of thousands of popular recipes) to answer a user’s query. An example of a supportable query is shown in Figure 1.

Figure 1: Example of an anomaly and trend in carbohydrate intake discovered by TSS and translated into RDF triples for the PHKG. A semantic reasoner reasons over the ADA guidelines, the Food KG and PHKG to answer the query.

Conclusions
Using the available set of ADA guideline OWL formalizations and the PHKG, a semantic reasoner can be used to infer whether the individual has been adhering to behaviors that are consistent with the guidelines. In any cases of violations, the semantic representations allow us to provide evidence-based recommendations based on their lifestyle and diabetes condition. Ongoing and future work is focused on (1) expanding the PHO to accommodate concepts important for comparing behaviors to ADA guidelines, (2) incorporating other knowledge resources such as the Healthy LifeStyle (HeLiS) ontology, which integrates knowledge related to food and activity.

Acknowledgement
This work is supported by IBM Research AI through the AI Horizons Network.

References
Promoting Use of Common Data Elements in Research Studies

Paul A. Harris, Ph.D., Rob Taylor, MA, Vaishali Jagtap, Douglas Conway, Stephany N Duda, Ph.D., Alex C. Cheng, Ph.D., MEM
Vanderbilt University Medical Center, Nashville, TN, USA

Introduction
The COVID-19 pandemic heightened interest in standardizing research data collection for observational studies and clinical trials. The U.S. National Library of Medicine (NLM) manages federally sanctioned common data element (CDE) sets in the NIH CDE Repository. While the NIH CDE Repository makes data elements available in a central location, the process of finding and transferring CDE details when creating study data collection instruments is cumbersome. REDCap (Research Electronic Data Capture) is a secure, web-based software platform designed to support data capture for research studies. REDCap is developed at Vanderbilt University Medical Center and widely used by research teams at more than 4900 partner institutions in 141 countries. We implemented a self-service workflow tool within REDCap to promote awareness and adoption of CDEs from the NIH CDE repository.

Methods
The REDCap CDE workflow tool uses an application programming interface (API) available from the NIH CDE Repository to search and retrieve metadata associated with NLM-curated CDEs. This API-driven approach enables real-time CDE content refresh and eliminates the need to download and store large amounts of CDE metadata within the REDCap application. After exploring CDE metadata content, our development team reviewed existing REDCap project creation workflow options and anchored the CDE discovery tool within the context of a larger workflow where research teams create case report forms or survey instruments for individual studies. In this process, researchers typically add a field to a data collection instrument by providing details about the desired data element (e.g. question label, type of input field, categorical options and codings, and validation requirements). The new CDE workflow tool simplifies this process by inviting end-users to “import from field bank” when adding a new data element. Once invoked as a workflow choice, research end-users are prompted to enter a keyword for real-time retrieval of associated CDEs within the NLM CDE Repository. Researchers can choose an option to retrieve information about CDEs curated by a specific institute or program such as NHLBI, NINDS, or RADx-UP. A keyword search triggers retrieval of one or more relevant CDEs. Then researchers are presented with a form-rendered version of the CDEs with relevant attribution details and options (e.g. contextual text labels) and are invited to assess fitness of use in their study. Upon selection of a specific CDE for inclusion, metadata details (e.g. question text, response option codes, response option text, and CDE ID number) are automatically transferred to the study data collection instrument and immediately available for research data collection and management.

Results
The REDCap CDE workflow tool was tested at Vanderbilt and disseminated to the REDCap Consortium in mid-January 2021 with REDCap version 10.7.0. Our decentralized model for REDCap dissemination does not allow for detailed evaluation of use from Consortium partner sites. However, we can access coarse usage details from institutions that choose to report statistics. In the two months post-release, we are aware of 357 partner institutions where the new CDE workflow tool is available to research teams. Researchers at 24 institutions have imported at least one CDE from the NIH CDE Repository into REDCap. Among these institutions, 189 CDEs have been imported. Researchers at Vanderbilt University Medical Center have imported 22 CDEs.

Discussion
We have created a user-friendly workflow designed to promote the use of CDEs by research teams using REDCap. This tool will not solve all problems associated with national and international adoption of CDEs, but we hypothesize that we can increase awareness and use of CDEs within the REDCap Consortium. In discussions with early adopters within the REDCap Consortium, we have received requests to consider content from additional CDE sources, including international CDE repositories. Future plans include evaluation and evolution of the CDE workflow tool with early adopters and potential expansion to include additional CDE repositories.

Funding: This work is supported by NIH/NLM contract #75N97019P00279
COVID-19 Dashboard: Visual Exploration of the Regional Pandemic Trend

Huan He, PhD 1, Liwei Wang, PhD 1, Andrew Wen, MS 1, Ming Huang PhD 1, Yanshan Wang, PhD 2, and Hongfang Liu, PhD 1

1Department of AI and Informatics Research, Mayo Clinic, Rochester, MN, USA
2Department of Health Information Management, University of Pittsburgh, Pittsburgh, PA, USA

Introduction: The fast spread of the coronavirus disease 2019 (COVID-19) has led to a worldwide pandemic and health crisis since December 2019. To address the urgent needs of tracking the regional trends of the COVID-19 outbreak, many organizations collected datasets from multiple sources and developed interactive dashboards to show the situation. However, due to the complexity of pandemic, it is challenging to explore and compare how the COVID-19 outbreak evolves in different regions. To address this challenge, we therefore present the COVID-19 dashboard to explore both the geographical and temporal patterns of the COVID-19 pandemic and update it daily.

System and Visual Design: To compare the regional pandemic difference intuitively, we proposed a new indicator, namely CrRW status, which is defined by combining the 7-day smoothed average case rate per 100k capita (Cr7d100k) and Cr7d100k ratio (i.e., the ratio of today Cr7d100k to 7 days ago) with the following empirical thresholds in the past seven days: the GREEN status, if Cr7d100k<15 and RW_Cr7d100k<1; the RED status, if Cr7d100k > 30, or if Cr7d100k > 15 and RW_Cr7d100k > 1.1; and the ORANGE status for everything else. To ensure that the dashboard could be updated continuously and automatically, we developed a data pipeline to collect data from public data sources (e.g., USAFacts, John Hopkins Coronavirus Resource Center, etc.) and calculate the indicators such as positive test rate, vaccinated percentage, as well as our proposed CrRW status for each region. We collaborated with domain experts from our clinic and developed the COVID-19 dashboard. As shown in Fig.1 (A), we designed a calendar-based heatmap to show the overall CrRW status changes. Users could click on the date cell to check the pandemic status of a specific date and the maps on different levels will be updated accordingly (Fig.1 (B)). We designed a novel trend chart to show how the pandemic changes overtime (Fig.1 (C)).

Discussion: As the COVID-19 pandemic situation changes, we could observe the trends by using this dashboard. We found that selecting appropriate indicators is important to capture the pandemic status accurately, especially when the pandemic varies from region to region. Although the outbreak has been significantly controlled by the non-pharmacological interventions and the massive vaccination, it is still not completely over. We will keep tracking the pandemic and adding new data such as new variants reports when dataset is available.

References
Overview and Descriptive Analysis of a New Ontology for Normalizing Section Types in Unstructured Clinical Notes

Paul M. Heider, PhD¹, Stéphane M. Meystre, MD, PhD¹
¹Medical University of South Carolina, Charleston, SC

Introduction: Context is critical to understanding the relevance of terms mentioned in unstructured clinical notes. Just as note types provide a gross sense of context, note sections refine that context by restricting the domain. Splitting notes into sections consists of two subtasks: finding boundaries and adding standardized labels. We present a new section type ontology to facilitate and standardize the work done in this second normalization subtask. Systems that address both subtasks can be mapped to our standard for easy cross-system comparison. Systems that only address the first subtask can benefit from more consistent corpus annotation. We provide descriptive statistics for two section-annotated corpora, including section labels for the first of these corpora which has not previously been normalized.

Methods: The core ontology was constructed by analyzing consistencies between other existing section type classification systems with the goal of comparing work across these different systems. These listings included: SecTag¹², Textractor¹, SOAP, LOINC, and an early prototype for processing clinical notes related to COVID-19 patient visits. We targeted a shallow hierarchy with depth introduced for section types that are as frequently addressed together as addressed individually (e.g., “Assessment/Plan”, “Assessment”, and “Plan”).

Our first corpus augmentation was on the i2b2 2014 Shared Task⁴, which consists of 1,304 de-identified clinical notes. Dai et al.⁵ released section split annotations (i.e., subtask 1) pre-annotated with machine-generated spans and corrected by the authors. We semi-automatically mapped these section header strings to our own section ontology types. Headers that directly matched an ontology label were mapped to that label. All other headers were manually reviewed for mapping by the first author with consultation with the second author. Our second corpus augmentation was to update the normalized section types in a corpus originally part of Textractor¹ development at the University of Utah. As the Textractor type system had already been mapped to the current ontology, this work consisted of updating the normalized string in the reference standard and porting it from Knowtator XML to brut and UIMA CAS XMI formats.

Results: The ontology consists of 29 top-level section types. Six of those concepts have one level of subtype. All top-level types and subtypes total to 50 unique section types. The Dai et al. corpus contains 13,133 total section header strings with 1,316 unique strings. Mapped to our new ontology, an average note has 10.1 headers with 8.5 unique section types. The Textractor corpus contains 2,920 total section header strings with 944 unique strings. Mapped to our new ontology, an average note has 34.0 headers with 13.2 unique section types.

The ontology structure, types, and mappings to prior section type systems is available via github: https://github.com/MUSC-TBIC. We cannot directly release our corpus augmentations due to restrictive rights. However, we have released (via github) a mapping script that can be applied on top of the Dai et al. annotations for the 2014 i2b2 corpus for individuals who have already gained access to this corpus through standard channels.

Conclusion: One outstanding issue with section normalization, that is also true of clinical note type normalization, is the common occurrence of clearly composed section types that are not indicative of the need for a supertype (e.g., in contrast to “past medical history” and “past surgical history” sitting under “past history”). The LOINC document ontology has taken a faceted approach by allowing multiple, orthogonal labels to be applied to a note (e.g., author’s role vs. type of service). More research is required to understand how pervasive and problematic these headers are.

Acknowledgements: This work is supported by PCORI (A20-0174-001).

References
Evaluating the Downstream Performance Impact of Various Common Off-the-Shelf Clinical NLP Components

Paul M. Heider, PhD1, Stéphane M. Meystre, MD, PhD1
1Medical University of South Carolina, Charleston, SC

Introduction: As the field of clinical natural language processing (NLP) matures, the available full system pipelines and individual customizable modules grow in pace. Researchers dutifully compare their systems to a subset of reasonable baselines. Often, however, performance metrics do not include variation introduced by upstream (pre-processing) modules or by their own contributions on downstream modules. We argue that part of our scientific due diligence is to facilitate and publish off-the-shelf protocols and scripts to cover these gaps in the literature. Prior studies1, 2 have documented system deployment and evaluation of the same task. We compare the downstream impact of system-critical NLP pre-processing modules. That is, how does the choice of tokenization and sentence boundary detection alter the (downstream) performance of a dictionary look-up module and a machine learning module?

Methods: To maintain consistency, we used a single (but configurable) Apache uimaFIT3 pipeline for testing. Output was evaluated using the ETUDE engine4, a freely available open-source evaluation tool for unstructured data and extractions. The pipeline, available via https://github.com/MUSC-TBIC, is described below. We focused on two upstream pre-processing tasks essential to most NLP work: sentence splitting and tokenization. For our dictionary look-up task, we used UMLS Metathesaurus-based dictionaries containing terms for symptoms commonly associated with COVID-19 and matched them using UIMA’s ConceptMapper. For our machine-learning task, we used the base version of a BiLSTM-CNN-CRF de-identification system trained on a MIMIC-III-based word embedding model.

Results: The top half of Table 1 presents scores for each tokenizer on symptom extraction, as the dictionary look-up was configured to ignore sentence boundaries. The bottom half of Table 1 presents scores for tokenizer and sentence splitter pairs. We found tokenization had a more significant impact on these particular downstream tasks than sentence boundary detection. Holding the tokenizer constant, different sentence boundary detection algorithms only impacted the F1-score by at most 5%. In contrast, holding the sentence splitter constant effected the F1-score by up to 56%.

Table 1. F1-Scores for Tokenizers on Symptom Extraction and Tokenizer/Sentence Splitter Pairs on De-identification

<table>
<thead>
<tr>
<th>Tokenizer</th>
<th>Sentence Splitter</th>
<th>Whitespace</th>
<th>Symbol</th>
<th>OpenNLP (Default)</th>
<th>OpenNLP (Aggressive)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0.2450</td>
<td>0.4278</td>
<td>0.4449</td>
<td>0.4463</td>
</tr>
<tr>
<td>cTAKES (Default)</td>
<td>Newline</td>
<td>0.2605</td>
<td>0.6876</td>
<td>0.4136</td>
<td>0.8355</td>
</tr>
<tr>
<td>cTAKES (BIO)</td>
<td>cTAKES (Default)</td>
<td>0.2699</td>
<td>0.6875</td>
<td>0.4114</td>
<td>0.8154</td>
</tr>
<tr>
<td>OpenNLP</td>
<td></td>
<td>0.2559</td>
<td>0.6570</td>
<td>0.4093</td>
<td>0.8094</td>
</tr>
</tbody>
</table>

Acknowledgements: This work was supported in part by PCORI (A20-0174-001) and by a research grant from NIGMS (R42GM116479). We thank the clinical domain experts who helped us annotate the 15 notes from MUSC’s telehealth system (Dr. Jihad Obeid, Dr. Hamilton Baker, Matthew Case and Michael Kopscik)

References

Integrating Medical and Dental Histories for Translational Science

Darren W. Henderson¹, Malini S. Kirakodu², James R. Aaron, MHA¹, Tamela J. Harper, MHA¹, Daniel R. Harris, PhD¹, Jeffery Talbert, PhD¹, Luciana M. Shaddox DDS, MS, PhD²

¹Center for Clinical and Translational Sciences, University of Kentucky, Lexington, KY 40506; ²College of Dentistry, University of Kentucky, Lexington, Kentucky 40506.

Abstract We present our efforts to integrate University of Kentucky College of Dentistry clinical data into our i2b2 warehouse. This linkage allows researchers to easily combine dental health history with other facets of the electronic medical record when predicting patient cohort sizes. By integrating these disparate clinical systems, we successfully link a meaningful subset of the dental patients (25%) and encounters (42%) into our existing clinical i2b2 data warehouse comprising both inpatient and ambulatory encounters in the UK Healthcare system.

Methods There is broadening evidence supporting the need to link oral health with overall health¹ and the integration of electronic dental data with other clinical data²,³. The University of Kentucky College of Dentistry averaged 44,072 patients per year from 2010-2020 (1.1 million visits in total). We extended i2b2 by adding an ontology matching the Current Dental Terminology (CDT) developed by the American Dental Association; local codes were added as children within CDT. This metadata development enabled the low-effort intersecting of dental observations with other clinical observations in i2b2. The dental data uses the multi-fact table approach supported by i2b2 and could be easily adopted by other institutions.

Discussion Using a deterministic method allowing fuzzy matching, we matched 25% of the available patients in the dentistry data to their general UK Healthcare record. Only 8% of patients strictly matched on social security number, name, and date of birth; allowing the social security number to be optional greatly increased the matches. Internal efforts are underway to improve the quality and completeness of patient records at the dental clinics which will further increase the linking success rate. With 25% of the population matched, we were able to link these dentistry patients to 17.2% of the UK Healthcare encounters, representing 13.6% of the overall clinical facts available in i2b2.

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th>Avg Per Patient</th>
<th>StDev</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distinct Patients in College of Dentistry Data</td>
<td>265,767</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patient Encounters in Dentistry Data</td>
<td>1,118,123</td>
<td>4</td>
<td>7.90</td>
</tr>
<tr>
<td>Patient Concepts in Dentistry Data</td>
<td>2,151,162</td>
<td>4</td>
<td>4.93</td>
</tr>
<tr>
<td>Dentistry Patients matched with UK Healthcare Population</td>
<td>62,131</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Matched Patient Encounters in Dentistry Data</td>
<td>470,907</td>
<td>7</td>
<td>11.47</td>
</tr>
<tr>
<td>Matched Patient Concepts in Dentistry Data</td>
<td>886,770</td>
<td>7</td>
<td>6.36</td>
</tr>
<tr>
<td>Matched Patient Encounters in UK Healthcare Data</td>
<td>3,116,697</td>
<td>7</td>
<td>50</td>
</tr>
<tr>
<td>Matched Patient Concepts in UK Healthcare Data</td>
<td>90,260,580</td>
<td>13.6%‡</td>
<td>231</td>
</tr>
<tr>
<td>Total UK Healthcare Encounters (1,290,410 patients)</td>
<td>18,097,392</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total UK Healthcare Facts</td>
<td>664,457,959</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Frequencies showing the general matching efficacy between the College of Dentistry data and UK Healthcare data. †Percent of total UK Healthcare Encounters (line 9).‡Percent of total UK Healthcare facts (line 10).

References


Development of a Health Equity Integration Dashboard (HEID) using Race Ethnicity and Language (REaL) Data in a Pediatric Hospital Setting: Lessons Learned in Visual Strategies and Key Data Elements

Raquel G. Hernandez MD MPH1,2,3, Paola Dees MD3, JoAnn DeRosa4, Fiorella Gonzalez4, Christine Hammerschmidt MA4, John Morrison MD PhD1,2,3, Benton Ng MD5, Ashish Shah MD MBA5, Luis Ahumada MSCS PhD1

1. Johns Hopkins All Children’s Hospital, Institute for Clinical and Translational Research, St. Petersburg, FL.
2. Johns Hopkins University School of Medicine, Department of Pediatrics, Baltimore, MD.
3. Johns Hopkins All Children’s Hospital, Department of Pediatric Medicine, St. Petersburg, FL.
4. Johns Hopkins All Children’s Hospital, Division of Quality and Safety, St. Petersburg, FL.
5. Johns Hopkins All Children’s Hospital, Heart Institute, St. Petersburg, FL.

Introduction: Efforts to promote health equity hinge on effective and visually intuitive methods of quantifying and monitoring clinical outcomes stratified by social determinants of health (SDOH). The development of health equity dashboards using health informatics strategies is a novel approach with favorable results in addressing disparities observed in early models. We describe our institution’s experience in developing and vetting a pediatric health equity integration dashboard (HEID) focused on addressing racial-ethnic disparities in hospital readmissions, length of stay and cardiac surgical mortality.

Methods: Johns Hopkins All Children’s Hospital (JHACH) is a 259 bed quaternary free-standing children’s hospital that is a safety-net for uninsured and underinsured children and serves a racial-ethnically and linguistically diverse population. Past QI efforts within the organization have leveraged EMR-based clinical outcomes focused on 30-day inpatient readmission and length of stay and cardiac surgical mortality. These outcomes have not previously been stratified by race, ethnicity or language (REaL). Composite REaL categories were created (e.g. Black-Non-Hispanic, Black Hispanic) allowing each clinical outcome to reflect a control group (lowest risk) and group of interest (highest risk). Visualization strategies for these data include: 1) use of empiric quarterly metrics vs. relative rates, 2) use of area under the curve (AUC) to compare rates, 3) limiting the use of traditional “traffic light” visuals and colors and promoting more objective color schemes to limit bias and inferences on observed disparities.

Results: Our Health Equity Integration Dashboard model (Figure 1) reflect three main strategies: 1) Display of Quarterly/Year metric rates by group of interest (dark blue) against comparison group (light blue). 2) Display of moving average rate into an AUC graphing approach to help display group disparities in parallel. 3) Linear depiction of the moving average difference between the group of interest and comparison group to help depict patterns of change and values close to zero indicating improvement in disparities. (Software: Tableau, version 2020.4.1; Data depicted on this chart are simulated and do not represent observed rates).

Figure 1: Health Equity Integration Dashboard

Discussion: Development of a health equity dashboard by leveraging existing datasets is an efficient and pragmatic approach to begin monitoring for health disparities. Creating robust racial-ethnic strategies and including key visualization and data elements such as those described may facilitate building of dashboards tailored to the needs of other hospitals and organizations. Inclusion of multi-disciplinary team perspectives is critical to promoting value and effectiveness in dashboard use. Such monitoring strategies will be critical to monitoring ongoing hospital QI interventions in real-time and promote timely and effective approaches towards health equity.
Expanding Exposure Notification Verification Code Issuance in WA State

Amanda Higgins, MSHI\textsuperscript{1,2,3}, Shannon O'Keefe, MSHI\textsuperscript{1}, Bryant T. Karras, MD\textsuperscript{1,2}
\textsuperscript{1}Washington State Department of Health, Olympia, WA, \textsuperscript{2}University of Washington, Seattle, WA, \textsuperscript{3}University of New England, Biddeford, ME

Abstract

Exposure notifications enhance contract tracing and case investigation efforts by decreasing the time in which known and unknown contacts are notified of possible exposure to COVID-19. This has the potential to reduce infections and save lives. In Washington State, additional methods of verification code issuance expand the functionality of the app by notifying more people as quickly as possible.

Introduction

In an unprecedented collaboration among competitors, Google and Apple partnered to develop Bluetooth based exposure notification technology originally proposed by researchers\textsuperscript{1} to assist the public health response to the COVID-19 pandemic. The Google | Apple protocol GAEN (Google-Apple Exposure Notifications) is a privacy-preserving tool that is available at no cost for public health authorities to enhance contact tracing and case investigation efforts by decreasing the time contacts are notified of possible exposure to SARS-CoV-2\textsuperscript{2}. GAEN-based technology works based on the exchange of random keys via low energy Bluetooth between phones that have exposure notifications enabled. When a user tests positive for COVID-19, they can anonymously consent to alert others they were near by entering a verification code, which triggers the upload of the random keys stored on their device to a national key server maintained by the Association of Public Health Laboratories (APHL). All phones with exposure notifications enabled periodically check the key server for a match for positive keys. If there is a match, the user will receive an alert that they may have been exposed and guidance from the public health authority on what to do next.

Methods

At the launch of WA Notify, verification codes were issued only during case interviews by case investigators and contact tracers to positive cases who attested that they were using WA Notify. This included an integration into the Case Management system (CREST) via an API to the verification code generator for Centralized State case investigators. Local health departments not using CREST manually issue codes to cases by logging in to the verification code generation portal. To reduce the time to issue codes, in January 2021, a mass SMS text issuance process was implemented to send verification codes to all phone numbers associated with a positive test reported to the Washington State Disease Reporting System (WDRS) the previous day.

Results

The addition of direct texting via mass SMS of the codes increased the use of Bluetooth Exposure Notification significantly. Central investigations issued 1477 codes successfully, Local Public Health interviews resulted in 1577 codes, and mass SMS resulted in 3070 claimed codes during January – March 2021.

Conclusion

The exposure notification app, WA Notify, is an integral tool along with masking, hand-hygiene, physical distancing, the recent widespread availability of vaccines to slow the spread of the virus. The increase in mass gatherings and travel as states re-open, breakthrough cases, and vaccine opt-out rates support the continued need for this valuable tool in the public health toolkit to reduce the spread of the virus and save lives.

References

Identifying Narrative Documentation of Clinician Concern about Patient Deterioration in Home Healthcare: A Text Mining Study

Mollie Hobensack, BSN, RN1,2, Jiyoun Song, PhD, RN1, Maryam Zolnoori, PhD1, Marietta Ojo, MPH1, Kathryn H. Bowles, PhD, RN2,4, Sena Chae, PhD, RN3, Erin Kennedy, BSN, RN4, Margaret V. McDonald, MSW2, Maxim Topaz, PhD, RN1,2

1Columbia University School of Nursing, New York, NY, USA; 2Center for Home Care Policy & Research, Visiting Nurse Service of New York, New York, NY, USA; 3University of Iowa School of Nursing, Iowa City, IA, USA; 4University of Pennsylvania School of Nursing, Philadelphia, PA, USA

Introduction

In the United States, home healthcare (HHC) agencies provide care to more than 3.4 million adults per year1. One of the key objectives of HHC is to identify early signs of patient clinical deterioration to activate prompt interventions reducing the risk of hospitalization and other adverse events2. There is evidence that HHC narrative documentation can inform identification of patients at risk for deterioration. This study explores an inclusive approach to identify patients at risk using status and situations described in clinical narrative notes.

Methods

A retrospective observational study with secondary data analysis was performed using HHC narrative notes (visit notes and care coordination notes) from one of the largest HHC agencies in the US from the years 2015 to 2017. To create the gold standard, expert annotators with a clinical background were asked to review a subset of random narrative notes and label each narrative note as either “concerning” or “not concerning.” Traditional Algorithms (Naive Bayes, Random Forest, and Support Vector Machines), Neural Network, and Transformer Algorithms (BERT and GPT-2) were applied to train and compare model performances, and identify important features.

Results

Of the 4,000 narrative notes, clinicians identified 20% (n=800) as “concerning.” Best performance across all models was achieved using the BERT algorithm as indicated in Table 1. Figure 2 visualizes the most informative features. Words that were associated with the highest entropy scores were “pain,” “lives,” “bed,” and “meds.”

<table>
<thead>
<tr>
<th>Model type</th>
<th>Precision</th>
<th>Recall</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random forest</td>
<td>0.73</td>
<td>0.64</td>
<td>0.67</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>0.90</td>
<td>0.50</td>
<td>0.69</td>
</tr>
<tr>
<td>SVM</td>
<td>0.75</td>
<td>0.69</td>
<td>0.71</td>
</tr>
<tr>
<td>CNN</td>
<td>0.72</td>
<td>0.62</td>
<td>0.66</td>
</tr>
<tr>
<td>BERT</td>
<td>0.76</td>
<td>0.72</td>
<td>0.73</td>
</tr>
<tr>
<td>GPT-2</td>
<td>0.74</td>
<td>0.61</td>
<td>0.63</td>
</tr>
</tbody>
</table>

Table 1. Machine learning predictive performance

Conclusion

This study is the first to build an NLP algorithm to predict HHC narrative notes as “concerning” or “not concerning” regarding adverse events using expert clinician judgment. The results of this study show the feasibility of creating algorithms that can help identify patients who are developing “concerning” clinical trends. Next steps are to embed the algorithm within an EHR system. If identified as “concerning,” these patients can then be prioritized for risk reduction interventions to support clinical decision making.

Funding: This work was supported by the AHRQ under the award number R01HS027742 and the NIH/NINR under the award number T32NR009356.

References

The Impact of Quality Feedback and Reporting on Hospital EHR Medication Safety Improvement

A Jay Holmgren, PhD MHI, David W Bates, MD Msc

1University of California, San Francisco, San Francisco, CA; 2Brigham & Women’s Hospital, Boston, MA; 3Harvard Medical School, Boston, MA

Introduction

One mechanism by which EHRs were expected to improve safety was the implementation of computerized provider order entry (CPOE), paired with clinical decision support (CDS). However, national CDS medication safety performance has been variable, with many hospitals not alerting on potentially harmful orders. One policy mechanism to improve quality is public quality reporting. We therefore sought to evaluate whether public quality reporting and feedback is a useful mechanism to incentivize hospital EHR quality improvement. We used national US hospital data from the Leapfrog CPOE Evaluation Tool to identify the effect of providing hospitals with feedback on their EHR medication safety performance. Leveraging a change in the scoring and feedback aspects of the evaluation in 2017, we used a regression discontinuity to assess whether hospitals that receive negative feedback regarding their safety performance improve more?

Methods

The CPOE Evaluation Tool is included as part of the Leapfrog Hospital Survey and is one of several process quality measures used by the Leapfrog Group in their evaluation and rating of US hospitals. The CPOE Evaluation uses simulated patients and orders to test whether a hospital’s production EHR system generates a clinical decision support order of any type when inputting an order that would cause an adverse drug event. Public scores displayed on the Leapfrog website are reported qualitatively, with hospitals scoring over 50% receiving “Full Demonstration of Safety Standards”, and hospitals from 30%–49.99% receiving “Substantial Demonstration”, and hospitals below 30% receiving “Completed Evaluation” The sample included hospitals who took the Leapfrog Hospital Survey, including the CPOE Evaluation Tool Evaluation, in both 2017 and 2018. This data was combined with American Hospital Association survey data to capture hospital characteristics. We employed a sharp regression discontinuity design between the top publicly reported score, “Full Demonstration” and the “Substantial Demonstration” score in 2017. Our dependent variable was hospital improvement from 2017 to 2018. We used the traditional OLS model, as well as a nonparametric method developed by Calonico et al.(2).

Results

The final sample consisted of 1,183 hospitals in a balanced panel. We identified a clear discontinuity in performance improvement at the 50% cutoff in 2017. In our regression models, hospitals on the cut-off receiving the “Substantial Demonstration” feedback was associated with improved performance in the subsequent year in both OLS (B=4.71, p<0.001) estimates and in our robust, non-parametric estimates (B=8.44, p=0.03) The effect was consistent both with hospital covariates, and robust to a wide array of robustness checks.

Discussion

Hospitals that received publicly reported negative feedback improved significantly more than those who receive positive feedback in the subsequent year. Despite this progress, there is considerable room for improvement especially in advanced decision support, with few hospitals receiving a perfect score. These results show that publicly reported feedback on specific dimensions of EHR-focused process quality such as medication safety can lead to improvement, even in the absence of direct financial incentives.

References

Validating the Value and the Usability of Real-Time Prescription Benefit Data

Megan A. Holsopple, PharmD, BCPS1, Kyle Campbell, PhD1, Anna Baer MS1, Mihir Parikh, MS1
1RxRevu Inc., Denver, CO

Introduction
Real-time prescription benefit (RTPB) is a relatively new technology being leveraged by prescribers and health systems. This technology equips prescribers with helpful insurance information medication (e.g., formulary coverage, alternative therapies, out-of-pocket costs) as they are electronically prescribing medications. Awareness of patient prescription affordability has been shown to help increase adherence, improve chronic disease state management, and reduce future healthcare costs through transparency of patient-specific pharmacy benefit coverage.1,2

Beginning January 2021, Medicare and Medicaid Services (CMS) and Medicare Advantage Part D plan employers, unions, or insurance carriers must implement an electronic RTPB tool capable of integrating with at least one electronic prescribing system or health record.3 While RTPB technology is transforming traditional prescribing workflows, the response and impact is still being established. The expectation is that through transparency of patient-specific pharmacy benefit coverage at the point-of-care, RTPB can provide prescribers with helpful insurance information to improve cost-of-care conversations, increase patient access to drug therapy, and optimize medication therapy selection when possible. This case report evaluates the impact of RTPB on prescriber behavior.

Methodology
A total of 138,488 ambulatory prescription orders were linked to RTPB transactions from 2 academic medical centers and associated community hospitals and ambulatory clinics. The primary objective was to assess prescriber response to payer alternatives (e.g., pharmacy and medication) and payer coverage alerts. Secondary objectives evaluated patient cost-savings as well as health system personnel time and costs associated with processing coverage alerts.

Key Results
Payer returned pharmacy or medication alternatives were displayed over 30% of the time with the most common alternatives including same drug class alternatives (e.g., HMG-CoA reductase inhibitors). Selection of a medication or pharmacy alternative ranged from 2% to 3%. Coverage alerts were presented to end-users up to 15% of the time with ‘prior authorization required’ and ‘refill too soon’ alerts being the most common. The coverage alert, “Refill too soon; Refillable after “x” date” appeared to be the most impactful alert resulting in provider response of “suspend order” 52% of the time.

Payer coverage information at the point-of-care also resulted in financial savings for patients with pharmacy or medication switches driving an annualized cost savings of $96,000. Brand to generic switches saved patients an average of $390/prescription annually and pharmacy switches from a retail to a mail order setting saving patients an average of $105/prescription annually.

Prior authorization avoidance where the prescriber selected a therapy alternative without a prior authorization resulted in up to 11 weeks of personnel time saved and upwards of $27,000 and $7,000 of pharmacist and pharmacy technician resource costs saved respectively.

Conclusion
Patient-specific prescription coverage information through RTPB enables data-driven prescribing decisions, behavior change, patient savings opportunities, and streamlines operational costs associated with processing medications that have coverage alerts.

References
Evaluation an Opioid Reversal Alert at an Academic Medical Center

Bernard M. Hsia, PharmD; Diana J. Schreier, PharmD, BCPS, MBA; Benjamin J. Anderson, PharmD, MPH; Jenna K. Lovely, PharmD, BCPS

Mayo Clinic - Rochester, MN

Purpose
The advent of computerized prescriber order entry systems (CPOE) has reduced medication errors by providing real-time alerts to prescribers. However, an excessive number of alerts can lead to alert fatigue resulting in alerts being bypassed 49-96% of the time. This project evaluated an opioid reversal alert in the institution’s electronic medical record (EHR) by determining how frequently it was being accepted or bypassed. The alert was created during our institution’s EHR implementation in 2018, and fires when an opioid order (scheduled doses or patient controlled analgesia) is placed with no active reversal agent on the patient’s profile. A naloxone order is recommended by the alert. This was the first analysis conducted since implementation to evaluate the effectiveness of this alert.

Methods
Data for a six-month period (08/27/2020 to 1/26/2021) were collected and analyzed using a self-reporting tool provided by our institution’s EHR vendor. Data regarding the number of actions taken, bypasses, departments with the most bypasses, and reasons for alert bypasses were reviewed. Based on the findings, a restriction was added to exclude alert firing on patients on end-of-life care. The alert was reassessed for an additional six-month period (4/1/2021 to 9/30/2021).

Results
The alert was bypassed 26.2% (n=5054) of the time. The hospice and palliative medicine department had the most bypassed alerts with 1183 (23.4% of all bypassed alerts). The patient with the most bypassed alerts (n=20) was on end of life care, which resulted in a modification to the alert. Preliminary 3-month data evaluating the modified alert are shown in Table 1.

Conclusion
In our analysis, the alert was firing and being bypassed an excessive amount on patients receiving end of life care. As a result, the alert was modified to prevent firing in patients with end-of-life care, thus reducing alert fatigue.

Table 1: Total Alerts Fired, Rate of actions and no actions taken

<table>
<thead>
<tr>
<th>Timeframe</th>
<th>Total Alerts</th>
<th>Actions Taken (%)</th>
<th>No Actions Taken (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>08/27/2020 – 1/26/2021</td>
<td>19,300</td>
<td>14,246 (73.8%)</td>
<td>5054 (26.2%)</td>
</tr>
<tr>
<td>4/01/2021 – 7/1/2021 (preliminary 3 months)</td>
<td>4,049</td>
<td>3,368 (83.2%)</td>
<td>679 (16.8%)</td>
</tr>
</tbody>
</table>

Table 2: Top Departments with highest number of bypassed alerts

<table>
<thead>
<tr>
<th>Department</th>
<th>Number of alerts bypassed (%) N=5,087</th>
<th>Number of alerts bypassed (%) N=679</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hospice and Palliative Medicine</td>
<td>1183 (23.4%)</td>
<td>41 (6%)</td>
</tr>
<tr>
<td>Hospitalist</td>
<td>867 (17.2%)</td>
<td>108 (15.9%)</td>
</tr>
<tr>
<td>Internal Medicine</td>
<td>767 (15.2%)</td>
<td>103 (15.2%)</td>
</tr>
</tbody>
</table>

References
COVID-19-Catalyzed and Informatics-Driven Evolution to a Learning Health System

Allen Hsiao, MD FAAP\textsuperscript{1,2}, Lisa Stump\textsuperscript{1}, & Nitu Kashyap, MD FAMIA\textsuperscript{1,2}

\textsuperscript{1}Yale New Haven Health, New Haven, CT; \textsuperscript{2}Yale School of Medicine, New Haven, CT

Abstract

We share the experience of how a health system and affiliated medical school rapidly evolved in response to the COVID-19 pandemic to become a true learning health system, driven by informatics and built upon an enterprise EHR and information technology foundation. Two separate organizations aligned to systematically integrate, evaluate, and disseminate new data and knowledge to the bedside to meet the rapidly evolving challenges facing clinicians and operational leaders.

Introduction

Yale New Haven Health is a five hospital (seven hospital campus) health system with over 2,500 beds predominantly along the coastline of Connecticut and one hospital in southern Rhode Island. It’s largest hospital, Yale New Haven Hospital, serves as the academic medical center for the separate but affiliated Yale School of Medicine. COVID-19, catalyzed tremendous changes that evolved the at times uneasy affiliate relationship into a true learning health system. We outline some of the key components of a learning health system and our COVID-related highlights\textsuperscript{1,2}:

**Culture & Leadership:** A centralized Hospital Incident Command Structure (HICS) was initiated bringing together senior leaders together in daily meetings to review COVID numbers, manage resources and expedite transfers. Daily huddles of the CMOs, egular employee Town Halls, and email communications were initiated. The culture shifted to embrace evidenced based medicine across all of the hospitals, adopting a single approach to clinical care informed by the infectious disease experts.

**Systematic Evidence:** Internal and external (CDC, WHO) data was regularly sought out and analyzed, informing clinical pathway CDS with regular updates. A single set of masking and PPE policies were adopted and modified by evidence, a massive call center stood up nearly overnight to support thousands of calls from the patients and community searching for information and advice. A single resuscitation and triage protocol was created and adopted for the entire system based upon the work of a multi-disciplinary group of clinicians, ethicists, and faculty.

**Informatics and Information Technology:** The CIO, CMIO, ACMIO, and other senior information technology leaders were quickly embedded into clinical and operational leadership; the enterprise EHR grew in its importance as a foundational system, serving as a contact tracing tool, data collection device, and CDS delivery system for COVID-19. Clinical pathways for COVID and MIS-C with complex decision support were created and integrated into the EHR. Telemedicine was adopted and embraced in record time, supporting >4,000 visits a day, up from 15 pre-COVID. Innovation and “failing fast” was embraced, leading to over 700 telemedicine carts were assembled by the IT team with off-the-shelf CPU, camera, and speaker parts.

**Data & Analytics:** The use of data and analytics was embraced as never before, with a system-wide real-time dashboard created built upon the EHR and enterprise datawarehouse, allowing operational and clinical leaders to monitor bed and ventilator availability across all of the hospitals in a single view. Two COVID predictive models were independently created by two different teams utilizing EHR, testing, and waste-water data to predict volumes. This information was used to drive bed planning, staffing, supply chain, and overall strategy. Other dashboards were quickly created to monitor telemedicine efforts, call center volume, and supply chain needs such as PPE and COVID-medication availability. Race/ethnicity data evaluated to identify vulnerable populations & for community messaging.

Conclusion

The COVID-19 pandemic has been a devastating challenge; however, it has also served as a catalyst to push the evolution of our separate health system and school of medicine into an informatics-driven learning health system. The change in culture, operational changes, embracing of evidence and data creates a LHS foundation that will serve our patients well and enable us to improve patient care not just for COVID but challenges to come.

References

1. [https://www.ahrq.gov/learning-health-systems/about.html](https://www.ahrq.gov/learning-health-systems/about.html)
The Combination of Graph Neuron Network and Multi-Task Learning Model for Cancer Drug Combination Prediction

Kanglin Hsieh, Ph.D., Yu-Chun, Hsu Ph.D., Xiaoqian Jiang Ph.D., Yejin Kim Ph.D.
Center for Secure Artificial Intelligence for Healthcare, School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX, USA

Introduction: Combination therapies have become a standard treatment strategy for complex diseases, including cancer, asthma, diabetes, and bacterial infections[1], by increasing therapeutic efficacy and reducing toxic side effects compared to standard monotherapies. High-throughput screening (HTS) instruments can profile and evaluate several markers for drug combination effects such as IC50, EC50 under cell level. However, the combinational searching space is proportional to the number of testing drugs and making HTS is not cost-effective for systematic screening. This challenge motivated us to apply A.I. for drug combinations. Currently, Knowledge graph (KG) is widely applied to drug repurposing by consideration of multiple modality information including drug targets, gene-gene interaction, side effects, and an indication of drugs. The growing trend of utilizing graphic neural networks (GNN) to learn and abstract the knowledge graph has demonstrated great potential. In this research, we highlight the potential of using GNN to build the embedding of KG and utilize the embedding for drug combination prediction.

Materials: We used several biological databases for building the biological network for drug combination, including Comparative Toxicogenomic Database (compound - compound target, compound - compound-related pathway, compound - disease, and disease - disease related pathway), Cancer Cell line Encyclopedia (Cancer Cell line - overexpression/suppression genes), and STRING (gene - gene). The Drugcomb portal is used to train and validate the drug combination results.

Methods: For determining high-confidence relationships between cell lines and gene expression, we adapted gene-wide z score which selected cell line - gene pair which z score is above 3 or below -3. We selected the protein-protein pairs which p-value is below 0.05. Other pairs have adapted the information provided by CTDbase. For training the graph embedding, we used GrapheSAGE, an inductive framework, which leverages node feature information to efficiently generate node embedding for previously unseen data. The model (Fig.1a) leverage five GrapheSAGE layers to learn five types of information including 1) drug - drug targets : 421,078 pairs, 2) gene - gene : 123,722 pairs, 3) cell line - gene : 179,715 pairs , 4) disease related information : 195,285 pairs, and 5) pathway-related information : 625,524 pairs. 90 % of pairs are used for training, and 10 % of pairs are used for testing. The node features come from Drug Repurposing Knowledge graph, the largest currently available knowledge graph embedding[2].

To test our hypothesis, we randomly selected 90 % of drug pairs within the same cell line for training, and 10 % for testing. Loewe score was applied to determine 1) antagonistic effect (Loewe <= -10), 2) additive effect (-10 < Loewe < 10), and 3) synergistic effect (Loewe >= 10). The prediction model adapted a multi-tasks deep neural network. The model is multiple task models(Fig.1b). Task one is the prediction of combination effects, and task two is the prediction of the Loewe score.

Results: We evaluate the model performance on the following tasks: 1) prediction of novel drug combinations, 2) prediction of novel drug combinations on new cell lines, and 3) prediction of novel drug combinations on novel drugs. For each task, we further examine the synergistic effect (classification task) and Loewe score (regression task). In this preliminary task, we set up the Loewe score over 10 as the threshold to calculate sensitivity, specificity, precision, recall, and AUC over each task. In general, the AUC of task one was 0.815, task two was 0.850, and task three was 0.675

Discussion: In this preliminary study, we only used the cell line and drug embedding information for prediction. This may limit the model performance. We will continually utilize more embedding information such as disease, pathway, and genes to improve the performance. We will also try to add other information such as gene regulation, gene expression data for further testing.

Conclusion: The GNN derived embedding has the potential to predict drug combination effects.
Disparity analysis of patient portal messaging use for COVID-19 in urban versus rural locality

Ming Huang1, Andrew Wen1, Huan He1, Liwei Wang1, Sijia Liu1, Yanshan Wang1, Nansu Zong1, Yue Yu1, Julie E Prigge2, Brian A Costello2, Nilay D Shah1, Henry H Ting1,3, Chyke Doubeni4, Jung-wei Fan1, Hongfang Liu1 and Christi A Patten5,6*

1Department of AI and Informatics, Mayo Clinic, Rochester, MN, USA; 2Center for Connected Care, Mayo Clinic, Rochester, MN, USA; 3Department of Cardiovascular Medicine, Mayo Clinic, Rochester, MN, USA; 4Department of Family Medicine, Mayo Clinic, Rochester, MN, USA; 5Center for Clinical and Translational Science, Community Engagement Program, Mayo Clinic, Rochester, MN, USA; 6Department of Psychiatry and Psychology, Mayo Clinic, Rochester, MN, USA

Introduction: The fast spread of the infectious coronavirus disease-2019 (COVID-19), caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has led to a worldwide pandemic with high morbidity and mortality rates since December 2019[1]. Rural populations are disproportionately affected by COVID-19 and vulnerable to its impacts due to limited resources (e.g., shortages of key pandemic specialists and underfunded hospitals) to respond to the pandemic. At the same time, COVID-19 has rapidly accelerated a change in health care delivery from in-person visits to virtual care [2]. The COVID-19 pandemic and the associated mitigation activities have caused patients to postpone or cancel non-acute medical encounters. Millions of clinic visits have transitioned to online platforms for continued care access [3]. Patient online services (patient portals) could serve as an attractive virtual platform and resource for patients from rural populations to seek support from their providers for COVID-19 and non-COVID related issues [4]. In this work, we propose to analyze the urban versus rural differences in portal messaging utilization for COVID-19 during the first wave of the pandemic.

Methods: We collected more than 1.4 million portal messages generated by patients in the midwestern US between February 1 and August 31, 2020. We then filtered the patient-generated messages (PGMs) associated with COVID-19 using relevant keywords (e.g., COVID-19, Pandemic, Coronavirus, SARS-CoV-2, and 2019-nCoV) and their synonyms and morphological variations. We performed descriptive statistics on the acquired PGM data set and examined the following four aspects of patient utilization of portal messaging for COVID-19 in the Midwest: (1) Patient counts. We calculated the patient counts in terms of five different roles of the patients: (A) the patient involved in EHR, (B) the patients involved in portal, (C) the portal users, (D) portal message senders, and (E) COVID-19 message senders. (2) Numbers of PGMs on COVID-19 and unique patient senders. We calculated the daily numbers of PGMs on COVID-19 and unique patient senders in the Midwest and its urban and rural areas. (3) Message use for COVID-19 related care. We analyzed the PGMs used for assessing COVID-19 symptoms and discussing COVID-19 diagnostic tests and results and care plan in the Midwest including urban and rural areas. Message use for other healthcare issues caused by the COVID-19 pandemic. We examined other healthcare issues related to COVID-19 reported in the messages to understand COVID-19 impacts on health services and patients.

Results: The urban to rural ratio of portal users, message senders, and COVID-19 message senders was 1.18, 1.31, and 1.79. The urban to rural ratio (1.69) of PGMs on COVID-19 was higher than that (1.43) of general PGMs. The urban-rural ratios of messaging were 1.72-1.85 for COVID-19 related care and 1.43-1.66 for other healthcare issues on COVID-19. The frequent senders of COVID-19 related messages were 40+ years old, female, married, white, and English language speakers. Compared to urban patients, the disproportion of message use was aggravated for sending messages on COVID-19 among rural patients.

Discussion: In the Midwestern health care setting, rural patients have significantly lower utilization of portal messaging than urban patients. The urban-rural disparity of portal messaging exacerbated in order of portal users, message senders, and COVID-19 message senders. Compared with urban patients, rural patients sent fewer messages for COVID-19 diagnosis and treatment but more messages on other COVID-19 related healthcare issues (e.g., isolation, anxiety). Results suggest opportunities for increasing equity in rural patient engagement in patient portals (in particular minority populations) for COVID-19.

Conclusions: The findings can provide us useful information about patient utilization difference of portal messaging to address COVID-19 crisis for improving patient-centered care.
Addition of auditable testing rational provides real-time information on hospital practice patterns

Michael S. Icardi MD, Larry A. Mole PharmD, Mark Holodniy
US Dept. of Veterans Affairs Public Health Laboratory, Palo Alto

Introduction
Early shortages of SARS-COV-2 virus testing reagents lead to rationing of testing resources. This necessitated redirection of supplies to address critical testing vs discretionary testing needs. Existing metadata, based on physician request for rapid testing and test utilization, was not enough to allow for supply re-allocation to cover regional outbreaks. We investigated if addition of metadata indicating if testing was ordered for Diagnostic, Screening or Monitoring would provide data enough to show ordering demand and triage test utilization.

Method
Test ordering and reporting data was standardized to require selection of testing category at the time of ordering into Diagnostic (Symptomatic patients/COVID-19 is suspected), Screening (Testing performed to enable scheduling of procedures or admissions) and Monitoring (testing performed as part of a series to allow early outbreak detection in long term care facilities). Ongoing data was collected starting from mid-November and broken down by region, site and type for comparison and outliers investigated. Comparison of December regional and facility data showed differences in overall testing demand were affected by local testing policy variation and outbreak status making total test utilization unsuitable for use to direct testing supplies toward critical vs discretionary test needs.

Figure 1. Standard test data contains only test results with no indication of rational for testing. Total test utilization vs positivity rates do not allow enough discrimination to determine critical test demand from discretionary testing.

Figure 2. Comparison of regions and member hospitals with A.) Non-outbreak, B.) Overwelked-outbreak C.) Well supplied-outbreak, with added metadata. Region B.) shows increased unclassified (referral) testing and decreased screening testing along with increased positivity in diagnostic testing compared to A. Region C. shows a similar diagnostic testing expansion and positivity increase as B. but is well compensated lacking expansion of referrals.. Of note Region A. appears to have similar testing demand as Region B but this is due primarily to discretionary testing and not critical need.

Conclusion:
Addition of auditable testing categories can provide real-time information on test utilization patterns not available in standard test metadata. This can be used to direct limited supply allocation to critical areas in emergent health responses.
Challenges and Opportunities of Information Technology Adoption by the Caregivers of Home Care Patients

Onimi Jademi, PhD, Abir Rahman, Bsc., and Güneş Koru, PhD
University of Maryland, Baltimore County, Baltimore, MD

Introduction: Home health care (also referred to as home care) is a Medicare reimbursed service where supportive care is provided by clinicians dispatched by Home Health Agencies (HHAs), i.e., skilled nurses, physical and occupational therapists, speech therapists etc. to support patients’ recovery and rehabilitation in their own homes. During home care episodes, the caregivers of home care patients are usually family members providing unpaid and non-skilled care and have to frequently interact with both home care clinicians and patients. Information technology (IT) presents tremendous potentials in terms of supporting caregivers in overcoming or managing some of their challenges, e.g., by reducing the complexity of caregiving tasks and by enabling communication mechanisms which can be used to alleviate emotional burden and reduce social isolation. No study has investigated what IT tools family caregivers adopt and what are the challenges and opportunities in doing so which can serve all stakeholders investing their resources to facilitate such adoption purposes. Therefore, this research 1. explored the IT tools adopted by the primary caregivers of home care patients, and 2. identified the challenges and opportunities related to caregivers’ IT adoption.

Methods: An interview-based qualitative approach was followed due to a lack of prior research and evidence related to the study objectives which provided rich contextual data. The participants for interviews were selected from the primary family caregivers of patients admitted by a home-health agency in the mid-Atlantic region of the United States. A total of fifteen caregivers were interviewed for this study as data saturation was achieved by this point. The Framework method was used to analyze the qualitative data resulting in a hierarchy of emerging themes and typologies. Member checks were employed to improve data accuracy by providing the transcripts of the interviews to the participants and asking whether they have any additions or corrections. Cohen’s Kappa analysis score of 76.92% was obtained showing substantial agreement range.

Results: (1) Eleven caregivers expressed requiring information regarding patient care, billing and insurance, lab results, diagnostic reports, and medication. Caregivers conveyed that often lack of care-giving information hindered their ability to provide effective care. Participants also expressed often finding information regarding billing and insurance was complex and were unsure what the insurance covered expenses were. Three participants expressed that access to a portal where they could get the test results and medication information would help reduce their need to visit physically or make phone calls. Ten participants who used IT regularly utilized publicly-available online search engines to obtain information about caregiving. Five participants reported using online social networks: four of them used Facebook as a forum to ask questions to verify information they found through online searches. Four participants were aware of the portal made available by the health system but they still prefer to use public search engines to search for medical information. Four participants used YouTube as a mean of learning about caregiving.

(2) The identified challenges are lack of awareness of IT solutions, unfavorable perception of IT, perceived complexity, information overload, lack of support, security concerns, and preferring face to face communication over IT. Ten caregivers reported being unaware of the patient portal already available to them. Another challenge was related to perceived complexity and lack of usefulness, where six participants felt that it was difficult for them to learn new tools due to complexity. One participant expressed a need for help and guidance in using patient portals. The participants also expressed insufficient guidance and training as the reason for lack of IT adoption. Six participants mentioned that the information overload experienced using search engines and social networks, and the effort needed to find relevant information among scattered pieces were overburdening. Three participants wished their HHAs were proactive and pointed them to better organized resources. The participants also brought up privacy, security, and trust issues as reasons for not adopting IT; three participants mentioned not using IT because of the fear of their data being stolen or leaked. Nine caregivers mentioned they prefer face-to-face communication over using IT to retrieve caregiving information. The opportunities include, availability of IT tools, availability of information resources, and improving digital literacy. Eleven caregivers used a device connected to the Internet regularly or frequently. Five caregivers were able to find necessary caregiving information using Google, YouTube, or Facebook. Despite the caregivers’ lack of awareness, a patient portal was available to all. Eleven of the caregiver responded having some form of digital literacy and are interested to learn new technology that can help them improve their caregiving. Two more caregivers responded they were willing to learn and use IT.

Conclusion: This study presents evidence about the variety of information needed by the caregivers and the room for improvement in using IT for better management and utilization of such information. Moving forward, HHAs and their partner healthcare organizations should more actively promote the IT solutions they make available to their patients and caregivers through their websites, social media channels, and various community events and activities. The solutions should have an intuitive user interface that is similar to other commonly applications that patients already use and free of jargon or require specific technical knowledge to operate. IT vendors should build related information privacy and security protections for caregivers into the design of their systems to alleviate caregivers security concerns.

References
Utility of the Observational Medical Outcomes Partnership Common Data Model for Representation of Non-Query Data Mappings in the Arden Syntax

Robert A. Jenders, MD, MS, FACP, FACMI, FHL7, FAMIA
Charles Drew University & University of California, Los Angeles, CA

Introduction

Arden Syntax is a formalism supervised by Health Level Seven International (HL7) for representation of procedural medical knowledge with the goal of facilitating sharing units of knowledge known as medical logic modules (MLMs). Some site-specific changes must occur for a knowledge base to be transferred from one site to another. Key to minimizing site-specific changes is the standardization of database linkages, which in turn requires identification of a standard data model, vocabularies and query syntax. While these linkages primarily are queries to retrieve data for processing in an MLM, they also include non-query data mappings such as event statements that are used to define MLM triggers and destination statements that route MLM output. This is sometimes known as the “curly braces problem” of Arden because of the syntactic construct used to enclose these site-specific references. While possible data models have been examined in HL7 for this purpose, none has yet been selected. The Fast Healthcare Interoperability Resources (FHIR) standard has been considered, but it continues to evolve rapidly. By contrast, the Observational Medical Outcomes Partnership Common Data Model (OMOP CDM) has evolved to v6.0 and is commonly used to aggregate data from disparate sources in registries and other health data repositories. Previous work demonstrated that the CDM is adequate to represent data query elements in MLMs. The present work extends this to assess the CDM’s utility as a standard data model for non-query data mappings in the Arden Syntax, which is important because of Arden’s use internationally in the systems of multiple vendors in organizations that would benefit from knowledge sharing.

Methods

A previously assembled, robust convenience sample of MLMs was examined. The MLMs were concatenated; regular expressions in a text editor were used to extract the non-query data mappings; and the data elements therein were identified. Each then was assessed to ascertain whether it could be represented using OMOP CDM v6.0.

Results

A total of 331 MLMs were pooled from 6 source CDS systems, including 19 from 3 vendor knowledge bases and 312 from 3 academic medical centers, encoded in Arden Syntax v2.1 but compatible with the latest v2.10. MLMs concerned mainly with lab tests were the most common (138/331 = 42%), followed by clinical assessment or classification (75/331 = 23%) and medication (45/331 = 14%). The remainder addressed administrative and miscellaneous topics. Each MLM contained at least one non-query data mapping. A total of 1457 explicit non-query data mappings were identified – 430 event statements (30% of all mappings) and 1027 destination statements (70%). These mappings were compared to the OMOP CDM to assess whether they could be represented. The most common event mappings were hospital admission (representable as a VISIT_OCCURRENCE), laboratory test result (as MEASUREMENT) and ordering of a medication (as DRUG_EXPOSURE). Although many MLMs had an unmapped, implicit output destination (usually the patient’s electronic chart), the plurality (429/1027 = 42%) of the explicit destination statements were email, while most of the rest were special project logs and intermediate interpretations intended as blackboard entries to be consumed by other MLMs. OMOP lacks an explicit concept for a message and its destination, but these could be represented as an OBSERVATION or NOTE, although, given the limitations on accepted concepts for some attributes in the CDM, there is no exact alignment between CDM and a destination. A limitation of the analysis is the absence of variables expressed with object notation in the MLM corpus which could have improved alignment between Arden Syntax data mappings and a standard data model.

Conclusions

OMOP CDM is adequate to represent the non-query data mappings found in a large, diverse corpus of Arden Syntax MLMs, although CDS output messages and their destinations lacked explicit supporting constructs in the CDM. However, because CDM can nicely represent query data elements and many non-query data mappings, consideration should be given for its use as a standard data model for the Arden Syntax. Future work includes creating a syntax for data references in Arden Syntax that incorporates CDM as part of developing Arden v3.0.
Discovering Adverse Events Induced by Drug Interactions in COVID-19 Patients

Eugene Jeong, MS1, Anna K. Person, MD1, Joanna L. Stollings, PharmD1, Lang Li, PhD2, You Chen, PhD1

1Vanderbilt University Medical Center, Nashville, TN; 2The Ohio State University, Columbus, OH

Abstract

Polypharmacy is highly associated with adverse clinical outcomes in COVID-19 patients. But few studies have investigated DDIs in COVID-19 patients. The aim of this study is to detect novel DDIs by a comprehensive analysis of FDA Adverse Event Reporting System data. We discovered 643 drug-drug combinations by using five DDI detection algorithms. 226 of them were novel. Thus, this study showed the potentiality of using five models for identifying potential drug-drug interactions in COVID-19 patients.

Introduction

COVID–19 patients with multiple comorbid illnesses are more likely to be using polypharmacy to treat their COVID–19 illness and comorbid conditions1. The more medications a patient is taking, the higher the risks of adverse events (AEs) induced by drug-drug interactions (DDIs). This study aims to identify AEs caused by potential DDIs in COVID-19 patients using five typical DDI-AE detection models.

Methods

Our study extracts 22,669 adverse event reports whose indications for use were COVID–19 in the FAERS database from January 2020 to December 2020. We call groups of those reports as the COVID-19 cohort. We also extracted the FAERS reports before the pandemic (from January 2004 to December 2019) to create the non-COVID-19 cohort. We pulled two groups of drugs for each of the cohorts: i) drugs used to treat COVID–19 (Drug 1), and ii) drugs used to treat illnesses other than COVID-19 (Drug 2). We detected the potential DDI-AEs using five typical DDI-AE models that rely solely on the information as described in a four-by-two contingency table2 (Table 1). The five algorithms are Ω shrinkage measure model, additive model, multiplicative model, combination risk score, and chi-square measure. The significances of DDI-AEs were assessed for COVID-19 and non-COVID-19 cohorts, respectively. We considered DDI-AEs confirmed by all five models as significant ones. Significant DDI-AEs appearing in both cohorts were not considered as DDI-AEs in COVID-19 patients, and thus they were excluded. We used DDI–AE and single drug-AE confirmed by existing works as gold standard datasets to validate whether we identified novel DDI-AEs in patients with COVID–19.

Table 1. Four-by-two contingency table.

<table>
<thead>
<tr>
<th></th>
<th>Target AE</th>
<th>All other AEs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neither Drug 1 nor Drug 2</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>Only Drug 1</td>
<td>c</td>
<td>d</td>
</tr>
<tr>
<td>Only Drug 2</td>
<td>e</td>
<td>f</td>
</tr>
<tr>
<td>Both Drug 1 and Drug 2</td>
<td>g</td>
<td>h</td>
</tr>
</tbody>
</table>

Results

The number of DDI-AEs were 2,083,200. Among all DDI-AEs investigated, the significances of the 643 were confirmed. The number of DDI-AEs identified in the golden standard datasets was 417, and the remaining 226 were novel ones.

Conclusion

This study augments the existing knowledge base for DDIs in patients with COVID-19 in that it discovered the potential DDIs from the spontaneous adverse event reports. Further genetic and chemical validation studies are needed to experimentally confirm the findings.

References

Automatic Construction of Biomedical Knowledge Graph from Covid-19 Literature

Chao Jiang, MS1, Victoria Ngo, PhD2, Richard Chapman, PhD3, Hongfang Liu1, Guoqian Jiang3, Nansu Zong, PhD4
1Department of Computer Science and Software Engineering, Auburn University, Auburn, AL; 2Department of Artificial Intelligence and Informatics Research, Mayo Clinic, Rochester, MN; 3University of California Davis Health, Sacramento, CA; 4Department of Computer Science, Oregon State University, Corvallis, OR

*Corresponding: Nansu Zong; zong.nansu@mayo.edu

Introduction
The open-resource literature hubs [1] have been created for tracking up-to-date, scientific information regarding the 2019 novel coronavirus (COVID-19), which forms knowledge bases to explore the explicit or implicit associations amongst diverse biomedical entities, further enabling researchers to answer clinical questions related to COVID-19. However, most knowledge graphs are constructed based on the co-occurred biomedical entities, which result in false-positive predictions in the applications (e.g., link prediction), as the co-occurrence in literature does not always represent the true biomedical associations. Here, we propose a method to leverage the variants of a Generative Adversarial Network (GAN) to automatically generate knowledge graphs from raw COVID-19 literature. The full version of the study is currently under the journal submission process.

Method and Results
Given a network G(V, E) where V stands for a set of vertices (i.e., biomedical concepts in the literature) and E represents the edges among two vertices (i.e., the co-occurrence of two concepts), edge E consists of a small number of known true associations L, and a large number of unknown true associations U. Our goal is to find true associations among U. We adapted the GAN variant (e.g., NetGAN[2] and CELL[3]). By sampling random walks from the raw graph which consists of unlabeled edges, the discriminator of GAN learns to distinguish the random walks sample. After training, the random walks sample from the generator will be used for filtering the unreal edge in the raw graph.

We generated two datasets for this study, 1) a synthetic dataset based on CORA-ML [4], and 2) a real dataset extracted from CORD-19-on-FHIR datasets [5]. For the synthetic dataset, it was formed with the positive associations labeled as 1 and sampled negative associations labeled as 0. For real data, we extracted a network from two annotated networks (LitCovid [6], PubTator [7]) via SPARQL query. Three types of nodes are used (Gene, Chemical, Mutation/Disease) with a total of 23,578 nodes and the corresponding 288,270 edges (i.e., 228,148 for chem-disease, 32,611 for gene-disease, and 27,511 for gene-chemical) were used to form a network in the experiment.

We designed three tasks to evaluate our work, in which we briefly show two here due to the page limit. We show four methods, “Base NetGAN”, “Base CELL”, “NetGAN”, and “CELL” for the test over the synthetic data. There is no big difference between the two methods when considering the base case (i.e., without using unlabeled associations) for “Base NetGAN” and “Base CELL”. However, when unlabeled information is taken into consideration, both methods achieved better performance compared with random classification. CELL overall outperforms NetGAN (0.828 for CELL vs 0.724 for NetGAN with synthetic data and 0.706 vs 0.549 with real data).

Discussion
In this study, we have proposed a method to automatically generate a knowledge graph by removing the co-occurred edges from biomedical literature. Despite the capability and stability of our investigated methods shown in this study, there are a few limitations. Firstly, we only used the limited labeled associations in our real dataset due to a limited number of resources. During the annotation, the annotators found some bias could be introduced when there is a vagueness for the concepts used in different biomedical literature. Secondly, despite achieving improvement with AUC in our real dataset compared with random classification, there is still a gap between the experimental results in a controlled environment and the adaptation of the proposed method for data curation in real scenarios. Performance improvement is still needed.

Acknowledgments
This work was supported by funding from the National Institutes of Health (NIH) NIGMS (K99GM135488).

References
A clinical decision support system (CDSS) ontology to facilitate portable vaccination CDSS rules: preliminary results

Xia Jing¹, PhD, Min Hua, PhD², Yang Gong, PhD³, James Cimino, MD⁴, David Robinson, MD⁵, Dean Sittig, PhD³, Paul Biondich, MD⁶, Adam Wright, PhD⁷, Christian Nøhr, PhD⁸, Tim Law, DO⁹, Arild Faxvaag, PhD¹⁰, Akash Indani, BS¹, Nina Hubig, PhD¹, Ronald Gimbel, PhD¹, Lior Rennert, PhD¹

¹Clemson University, Clemson, SC, USA; ²George Mason University, Fairfax, VA, USA; ³University of Texas Health Sciences Center at Houston, Houston, TX, USA; ⁴University of Alabama at Birmingham, Birmingham, AL, USA; ⁵Independent Consultant, UK; ⁶Indiana University, Indianapolis, IN, USA; ⁷Vanderbilt University, Nashville, TN, USA; ⁸University of Southern Denmark, Denmark; ⁹High Mark BCBS, PA, USA; ¹⁰Norwegian University of Science and Technology, Norway

Introduction
Health IT has experienced large and rapid growth in the last several decades. In the United States, electronic health record (EHR) adoption rates among office-based physicians increased from 20.8% in 2004 to 85.9% in 2017, more than a four-fold increase in 14 years. By 2015, CDSS use in office-based primary care settings increased to a range between 68.5% and 100%, depending on the type of CDSS[1]. Nevertheless, the interoperable patient record is still not a reality in routine healthcare services after over a decade, as seen in the case of e-patient Dave. Our current effort is to build a CDSS ontology to facilitate interoperable CDSS rules by using CDC-recommended vaccination schedules as examples.

Methods
Our first step is to construct the ontology. The manual curation process includes at least four lines of effort: (1) the investigation team brainstorm, engages in iterative discussion, and provides recommendations based on their experience. (2) The investigation team conducts a manual review of the automatic method results. (3) The investigation team conducts thought experiments in regard to scenarios for two vaccines (MMR and HPV) in terms of the following key components: development, management, maintenance, and operation of CDSS rules. AHRQ CDS implementation guides and the existing CDSS taxonomy[2] are used to guide the thought processes, and the team records the detailed processes and extracts concepts that should be included in the ontology. (4) The investigation team conducts a manual review of and selects/adopts key concepts from the relevant ontologies, NIH common data elements, book chapters from CDSS (particularly Chapters 3, 15-20, 28 and 29)[3], AHRQ CDS knowledge artifacts, HL7 CDS standards and implementation guidelines, ONC US Core Data for Interoperability (USCDI), and CDC CDS for Immunization & Immunization Information Systems as well as other resources. The automatic approach focuses on using natural language processing approaches to identify relevant entities from the selected literature, from which the investigation team will decide whether or not to include and how to classify the entities.

Results
The basic metrics of the current version of the ontology are as follows: 160 classes, 97 properties, and 547 axioms. The broader categories and main structures of the ontology include activities (events), components and behaviors of CDSS, all of which provide building blocks for CDSS rules in the next stage.

Discussion
Our effort is ongoing, and the ontology will be published via Bioportal. Ideally, if we can use some production-level documents/resources as sources for the CDSS ontology, that would be helpful. However, such documents are not publicly available. The artifacts from AHRO and implementation guides and the standards from HL7 that we used as sources can substitute for these documents. The CDSS ontology has the potential to enable CDSS rules’ reuse, sharing, and maintenance, which will directly affect whether CDSS is using the most updated rules. The routinely updated CDSS rules, especially in under-resourced settings, can improve adherence to clinical guidelines and enable CDSS to play a critical role in clinical quality assurance.

Acknowledgment
The work is supported by the National Institute Of General Medical Sciences of the National Institutes of Health under Award Number R01GM138589.

References
Patient-Specific Modeling with Lazy Random Forest (LazyRF)

Adriana Johnson MS1, Gregory F Cooper MD PhD1, Shyam Visweswaran MD PhD1
1Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh PA

Introduction. The goal of precision medicine is to provide high-quality, individualized care for each patient, and predictive models can be useful tools in achieving that goal. However, standard machine learning methods may fail to capture informative patterns that are only present in small subgroups of the population. Personalized machine learning methods optimize a predictive model for each individual patient and may provide more informative predictions for each member of the population. Personalized decision paths (PDPs) are a type of personalized method based on decision trees introduced by Friedman in the form of the LazyDT1. Several types of PDPs have been assessed in the biomedical domain, often demonstrating superior performance to population methods2. However, little has been published on ensembles of PDPs. We introduce a novel ensemble method of PDPs based on random forests called Lazy Random Forest (LazyRF) and assess its performance on six clinical and genomic datasets for patient-specific prediction.

Methodology. LazyRF consists of bagged randomized PDPs. Using bootstrapped datasets, PDPs are trained on random subsets of the features present in the patient of interest, and features that optimize an information gain scoring criterion are included in the model. The resulting patient-specific ensemble of PDPs can be used for prediction by averaging the predictions from each path in the forest. We compared the predictive performance of single, non-randomized PDPs, bagged non-randomized PDPs, and our novel LazyRF method on six clinical and genomic datasets in terms of area under the receiver operating characteristic curve (AUROC). We also compared the performance of our LazyRF to a standard random forest of decision trees in terms of AUROC and model complexity as measured by mean predictive path length. We compared AUROC pairwise using DeLong’s test on each dataset, and compared AUROC and mean path lengths across all six datasets using the Wilcoxon Signed Rank Test.

Results. Table 1 contains the results of our experiments. Our LazyRF had statistically significantly higher AUROCs than the single PDPs for 5/6 datasets (indicated in bold), statistically significantly higher AUROCs than the bagged PDPs for 4/6 datasets (indicated in italics), and statistically significantly higher AUROCs than standard random forest for 1/6 datasets (indicated in red) as measured by DeLong’s test at the 0.05 level. LazyRF had statistically significantly higher average AUROCs overall than the single PDPs as measured by the Wilcoxon Signed Rank Test at the 0.05 level. The standard random forest had a statistically significantly higher AUROC than LazyRF for one dataset when compared via DeLong’s test (indicated in red) at the 0.05 level. When average performance was compared via Wilcoxon Signed Rank Test, the random forest and LazyRF did not have statistically significantly different AUROCs, but LazyRF had statistically significantly shorter mean path lengths than the random forest at the 0.05 level.

Table 1. Results for PDP, bagged PDP, LazyRF, and random forest

<table>
<thead>
<tr>
<th>Model</th>
<th>PDP AUROC</th>
<th>Bagged PDP AUROC</th>
<th>LazyRF AUROC</th>
<th>LazyRF Mean Path Length</th>
<th>Random Forest AUROC</th>
<th>Random Forest Mean Path Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chronic-pancreatitis</td>
<td>0.736</td>
<td>0.813</td>
<td>0.824</td>
<td>3.73</td>
<td>0.791</td>
<td>5.09</td>
</tr>
<tr>
<td>Pneumonia</td>
<td>0.512</td>
<td>0.544</td>
<td>0.542</td>
<td>4.18</td>
<td>0.743</td>
<td>5.66</td>
</tr>
<tr>
<td>Sepsis-D</td>
<td>0.551</td>
<td>0.696</td>
<td>0.840</td>
<td>3.00</td>
<td>0.852</td>
<td>5.82</td>
</tr>
<tr>
<td>Sepsis-S</td>
<td>0.641</td>
<td>0.718</td>
<td>0.768</td>
<td>3.41</td>
<td>0.753</td>
<td>5.95</td>
</tr>
<tr>
<td>Heart Failure-D</td>
<td>0.545</td>
<td>0.580</td>
<td>0.724</td>
<td>4.89</td>
<td>0.643</td>
<td>8.34</td>
</tr>
<tr>
<td>Heart Failure-C</td>
<td>0.515</td>
<td>0.567</td>
<td>0.711</td>
<td>6.08</td>
<td>0.698</td>
<td>9.86</td>
</tr>
</tbody>
</table>

Conclusions. Our results show that a novel personalized random forest method, LazyRF, achieves higher predictive performance than simpler personalized methods. Compared to the traditional random forest method, LazyRF has comparable predictive performance, but produces simpler models, which may be easier to explain. LazyRF may therefore be useful for providing accurate, personalized predictions in support of precision medicine.

Acknowledgements. We would like to acknowledge NLM grant T15LM007059 for supporting this work.

References
Clinical Research Coordinators’ Perceptions of Electronic Data Capture Systems Adopted in Clinical Research

Michelle E. Johnson¹, Fei Yu¹, PhD
¹University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Introduction
Despite their critical role in the success of clinical research studies, the needs of clinical research coordinators (CRCs) and study nurses are rarely addressed in clinical research informatics research. Additionally, few studies have evaluated CRC perceptions towards electronic data capture (EDC) systems or whether such perceptions reveal critical systems design challenges.¹ To address this gap, this study aims to investigate user experiences (UX) and perceptions towards common EDC systems, identify unmet information needs of CRCs, and address potential barriers to EDC technology acceptance by those who utilize these systems the most.

Methods
After the Institutional Review Board approved this study (IRB# 20-2620), semi-structured interviews were held with 15 CRCs working across ten study sites throughout the U.S. Interviews were audio-recorded, manually transcribed, and coded in NVivo 12² using an inductive, data-driven approach.³ Following a reflexive thematic analysis approach, recurring patterns were identified from transcribed data and collated into major themes. Then, themes were contextualized according to constructs within the Technology Acceptance Model (TAM) and UX models.

Results and Discussion
We identified 5 major themes related to EDC acceptance and CRC perceptions (Table 1). Within these themes, we confirmed ease of use and perceived usefulness as important facilitators of EDC technology acceptance. Identified themes will be used to construct a set of guidelines and recommendations for improving the design and overall user experience of EDC systems for CRCs.

Table 1. Themes identified from the qualitative interviews using thematic analysis.

<table>
<thead>
<tr>
<th>Preliminary Themes</th>
<th>Context</th>
</tr>
</thead>
<tbody>
<tr>
<td>Impact of the COVID-19 pandemic on data collection</td>
<td>Increased use of electronic source documentation and remote monitoring. Shift of workload towards in-patient studies with more repetitive data entry</td>
</tr>
<tr>
<td>Discrepancies between source and EDC</td>
<td>EDC forms not seen by CRCs until “first patient”; disparate and sometimes conflicting information needs between sponsor/site.</td>
</tr>
<tr>
<td>Queries both facilitate and challenge CRC workflows</td>
<td>Well-designed auto-generated queries can facilitate transparency, efficiency, and accuracy of data entry, while poorly designed auto-generated queries result in additional data entry burden.</td>
</tr>
<tr>
<td>Usability concerns pertaining to CRC-system interaction</td>
<td>Inability to save and return to incomplete forms; character limits; excessive scrolling and clicks; low ease of learning; lack of navigational shortcuts</td>
</tr>
<tr>
<td>Information fragmentation</td>
<td>CRCs resort to developing personal information systems to address study needs. Decreased task efficiency when moving between systems.</td>
</tr>
</tbody>
</table>

Conclusion
Current EDC systems do not meet the needs of CRCs. Our findings will help to inform the design and improvement of future EDC technologies.

References
2. QSR International Pty Ltd. NVivo 12. 2020.
Exploring Effectiveness of Domain and Task-Adaptive Pretraining for Clinical Information Extraction

Venkata Joopudi, MS, Ananya Poddar, MS, Bharath Dandala, PhD, Ching-Huei Tsou, PhD
IBM TJ Watson Research Center, Yorktown Heights, NY

Introduction
Pretrained transformer-based language models have demonstrated state-of-the-art performance on a wide variety of clinical natural language processing (NLP) tasks. Recent advances in adaptive pretraining methodologies have been mostly limited to the general domain. In this study, we apply this novel methodology to the clinical domain to establish best practices, and to assess if further adaptation to the task’s unlabeled data (task-adaptive pretraining (TAPT)), can help boost the performance of clinical NLP tasks, such as note-section recognition and clinical summary categorization.

Materials and Methods
We use expert-labeled sections on 1,303 summary sentences over 63,016 sentences from the i2b2 corpora for the note-section recognition task, and 84,424 summary spans and 63,594 summary sentences for the span-based* and sentence-based* clinical summary categorization tasks respectively. We use a BERTBASE* cased model to establish the baseline and utilize Clinical BioBERT (pretrained on biomedical, and further on clinical corpora) as our domain-adaptive pretraining (DAPT) model. For each task, we further pretrain the DAPT model using task-specific training data (DAPT + TAPT), with an unsupervised masked language modeling objective. In this process, we artificially augment the task-specific dataset by randomly masking different words (with a 0.15 masking probability) across 100 epochs.

During fine-tuning, for the note-section and sentence-based summary categorization tasks, we pass the final layer’s [CLS] token representation to a task-specific feedforward layer for classification. For the span-based summary categorization task, we pass the final layer’s token representations \{T_1, ..., T_N\} to a CRF layer for token classification.

Results and Conclusion

Table 1. Fine-tuning results for three clinical NLP tasks across different pretraining methodologies.

<table>
<thead>
<tr>
<th>Task</th>
<th># of Classes</th>
<th>Baseline</th>
<th>DAPT</th>
<th>DAPT + TAPT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Precision</td>
<td>Recall</td>
<td>Micro-F_1</td>
</tr>
<tr>
<td>i2b2 note-section recognition</td>
<td>14</td>
<td>0.6612</td>
<td>0.7374</td>
<td>0.6972</td>
</tr>
<tr>
<td>Span-based clinical summary</td>
<td>9</td>
<td>0.7711</td>
<td>0.7286</td>
<td>0.7492</td>
</tr>
<tr>
<td>categorization</td>
<td></td>
<td>0.7967</td>
<td>0.6861</td>
<td>0.7377</td>
</tr>
</tbody>
</table>

As shown in Table 1, our results confirm the well-known observation that DAPT improves performance on multiple clinical NLP tasks. Additionally, further pretraining of a DAPT model on a task-specific dataset (DAPT + TAPT) consistently improves performance on multiple tasks in the clinical domain. Through this study, we establish a better practice to leverage the limited task-specific dataset more efficiently to obtain considerable performance improvements, particularly in a clinical setting where manual curation of ground truth is expensive. Future work includes exploring data augmentation techniques that leverage semantic information (such as synonyms and hypernyms) available in knowledge bases to further expand the task-specific pretraining dataset.

References

* We normalize the span-based clinical summary dataset into a sentence-based dataset to account for the inter-annotator differences for the marked spans. We perform this normalization by first segmenting each note into sentences, and then expanding the human-annotated summary span label to the entire sentence.
Perceptions on Granular Data Sharing in Behavioral Health Populations

Tina Kaing, BS1, George Karway, MS1, Julia Ivanova, MA1, Anita Murcko, MD1, Michael Franczak, PhD2, Mary-Jo Whitfield, MSW3, Adela Grando, PhD1

1College of Health Solutions, Arizona State University, Scottsdale, AZ; 2COPA Health, Phoenix, AZ; 3Jewish Family and Children’s Services, Phoenix, AZ

Introduction: While studies indicate patients desire more control over their health data, little is known about data sharing preferences among behavioral health patients with behavioral health conditions1. The Substance Abuse and Mental Health Services Administration (SAMHSA) has developed a tool, Consent2Share, to give patients granular control over how their medical records are shared2; however, Consent2Share has not been evaluated with patients.

Objectives: Conduct a prospective, phone-based study within adult Spanish and English-speaking populations with general mental illness (GMI) and serious mental illness (SMI) to evaluate whether (1) patients desire granular sharing options offered by Consent2Share, (2) data recipient, data type, and purpose of use influence granular data sharing choices, and (3) supported granular data choices (domestic violence, genetic information, substance use, sexual and reproductive health, and mental health) satisfy patients’ data privacy needs.

Methods: Participants were recruited by phone among two integrated care clinics. Participants completed a demographics survey, answered questions related to three data sharing scenarios (sharing with primary care provider (PCP) within a shared facility, sharing with PCP at an outside facility, and sharing with a behavioral health provider (BHP) at an outside facility) supported by Consent2Share and completed a survey about the study and the supported granular data choices. Descriptive statistics and correlations were assessed.

Results: 200 participants were recruited (80 with SMI and 120 with GMI; 5 Spanish speakers and 195 English speakers). A majority (87%) of respondents indicated that knowing who or what facility would be using their data (either for care or research) improved overall levels of comfort with sharing. Participants were generally more willing to share within a shared facility as opposed to an outside facility (Figure 1), with participants diagnosed with GMI sharing more on average than those with SMI. Among the granular data categories provided, most (83%) participants indicated that supported granular data choices were sufficient to fully capture sharing preferences.

![Figure 1. Willingness to share between GMI and SMI patients, based on data recipient scenarios.](image)

Conclusions: Findings suggest a strong desire for mechanisms to support granular information sharing in current consent processes. Provided data sharing choices accommodated patient data privacy needs. Outcomes of this study will guide the development of future technologies and initiatives towards granular data sharing.

References

Introduction: Increasingly data are leveraged to promote health outcomes, and practitioners are turning to data from different economic sectors to address social determinants of health. Unfortunately, the US does not have a comprehensive data protection law; instead there is a patchwork of laws. The variation in data protection laws is particularly vexing for efforts to promote data sharing to promote population health. In this review we have aggregated and screened through the publicly available resources that may help public health officials and practitioners navigate the data sharing landscape.

Materials and Methods: We used a general-purpose search engine (www.google.com) to identify the resources published between 2010 and 2020. We developed a complex search pattern yielding 75 individual searches, where each search included a common string pattern: [common search stem] + [sector] + [data protection term]. The common search stem was: ("data sharing" OR "data use" OR "information sharing" OR "information use") + ("law" OR "regulation" OR "legislation" OR "statute"). We identified search terms to target a total of 24 different sectors and an additional set of searches were executed without a specified sector (24 sectors and 1 overall). The common search stem and the sector search terms were executed a total of three times, each with a different data protection term: “privacy”, “confidentiality” or “consent” (in that order). This search pattern yielded a total of 75 individual searches and first 50 hits for each search were saved. We scored each resource on a scale of 1 to 4 (lowest: 1, highest: 4) in terms of their depth of legal issues discussion, depth of data sharing discussion and value for addressing legal barriers.

Results: Our sector specific searches provided a total of 3710 hits, out of which 989 were duplicates. After removing the duplicates, remaining 2721 unique URLs were subjected to scoping screening. Full-text of 321 in-scope resources were reviewed by two reviewers, out of which 164 were selected for final coding. Among the resources selected for coding, an upward trend in number between the years 2010 to 2018 was observed. A total of 22 sectors (including 1 ‘other’ category) were covered in the included resources. Among them, clinical healthcare was the most prominent sector with a mention in 76 websites/files. The reviewed resources covered a total of 66 laws in different degrees but plurality of them focused on only a handful of laws (only 18 laws were discussed in more than 4 resources). HIPAA, FERPA and 42 CFR Part 2 were the top three federal laws discussed in the resources with higher depth of legal discussion (i.e. scored 3 or 4 out of 4). Among the 164 full-text resources coded, around 14% (n=23) discussed legal concepts and issues (e.g., consent, legal agreements) but did not expressly discuss a specific law, around 70% (n=114) discussed >1 laws and around 85% (n=139) discussed >1 sectors. In terms of depth of legal issues discussion, depth of data sharing discussion and value for addressing legal barriers, the number of resources with a score of 4 was 48, 39 and 48, respectively.

Discussion: This is the first systematic study of publicly available resources on legal data sharing issues. Our findings suggest that good resources are difficult to find. Practitioners trying to find pertinent resources will have to sift through voluminous search results that are not useful to identify, understand, and address legal barriers to data sharing. A relatively small proportion (1.76%, n=48) of our 2721 unique search results were scored as having the greatest depth of legal discussion. Consequently, the difficulty of finding quality resources likely amplifies the perception of legal data sharing barriers among practitioners.

Conclusion: Our findings describe the existing landscape of publicly available resources addressing legal data sharing issue and can help identify future needs.

Acknowledgements
This work was supported by the Network for Public Health Law, Data Across Sectors for Health, and the Texas A&M Population Informatics Lab.

References
Expanding Access and Continuity of care during COVID Crisis, in Safety Net Hospital
Sentaye Kassa, Brett Moran, Parkland Health and Hospital System

Introduction: When COVID-19 started to spread and affect patient care in March 2020, continuity of care and access to services were challenges for all our patients, particularly those living in undeserved communities. Telemedicine has experienced marked growth both in the US and globally over the past few years with exponential adoption noted with the COVID-19 pandemic [1]. While there has been promotion and advocacy of telemedicine by the government, reimbursement by CMS has been limited to only telemedicine which includes video, with special waiver for audio-only visits allowable only during the pandemic [2]. Uptake of audio-video telemedicine has been limited in some health systems due to budgetary issues as well as problems surrounding the “digital divide”, particularly impacting safety net systems. In Dallas, over 40% of households are without fixed internet [3], leading the top 10 cities in the US with largest digital divide [4]

Methods: During the beginning of the COVID-19 crisis in March 2020, we had to think outside the box to continue caring for our patients. We started training physicians and other hospital staff through WebEx, Pathway Modules, and a hands-on recorded didactic tutorial, then expanded virtual care to all specialty departments. This initiative was widely implemented across the hospital system within two to three weeks. This led to uninterrupted continuity of care for our vulnerable patients, protecting them and the staff from COVID. Most visits were conducted via telephone, and no personal information was compromised.

Results: Pre-COVID, we had a hospital-wide campaign to improve MyChart sign up rates, which increased MyChart utilization from 21% to 30%+. This facilitated the implementation of virtual care. The average Virtual Visit from April 2019 to Feb 2020 was about 5000 per month, but after the wide implementation of Virtual Visit the numbers from March 2020 to December 2020 increased to about 40,000 per month, an 800% increase. When it comes to show rate, the data shows that from Jan 2019 to Dec 2020, Virtual Visit show rate averaged 83.5 % while the face-to-face show rate average was 77.39 %. A survey was distributed from Sept 29, 2020, to October 23, 2020, via email to 774 physicians and APPs from 65 departments. 440 of them responded. Press Ganey patient portal data was reviewed regarding patient Audio-only virtual care experiences.

Conclusions: During COVID-19, through rapid expansion of Audio-only Virtual Care, we were able to reach out to our undeserved patients, overcoming the health disparity in primary and specialty care. When our homeless patients were moved to a hotel and isolated, the only way to get to them was through Virtual Care. We implemented Audio-only Virtual Visits. We improved access, sustained continuity of care, and engaged our patients, and as a result we observed better show rate, staff, and patient satisfaction. We gathered data by survey, Press Ganey patient portal, Slicer Dicer, and Dashboard. We can attest that Audio-only Virtual Visits can be successful in diverse undeserved populations.

Bibliography


Systematic identification of ACE2 expression modulators reveals cardiomyopathy as a risk factor for mortality in COVID-19 patients

Navchetan Kaur1, Boris Oskotsky1, Atul J. Butte1, and Zicheng Hu1,*
1 Bakar Computational Health Sciences Institute, University of California, San Francisco, San Francisco, CA. *Correspondence may be addressed to zicheng.hu@ucsf.edu

Introduction: Angiotensin-converting enzyme 2 (ACE2) is the cell-entry receptor for SARS-CoV-2[1]. The binding between ACE2 and spike (S) protein of SARS-COV-2 initiates the viral entry into target cells. ACE2 plays key roles in both the transmission and pathogenesis of SARS-CoV-2. To comprehensively profile the expression patterns of ACE2, we not only need to characterize its expression in healthy tissues but also identify diseases, drugs, and genetic perturbations that modulate ACE2 expression.

Methods: We developed a computational framework named GENEVA (Gene Expression Variance Analysis) to identify the most relevant datasets for visualization and detailed manual analysis. GENEVA prioritizes the datasets that have a large variance of ACE2 expression. At the same time, GENEVA controls for the overall heterogeneity of the samples to prioritize datasets in which ACE2 is specifically modulated by experimental conditions rather than due to tissue type differences. In addition, GENEVA embeds the meta-data into numerical space and prioritizes datasets with high correlations between ACE2 expression and the metadata.

Results: We applied the GENEVA tool to 9,124 publicly available RNA-seq datasets. We identified multiple drugs and diseases that modulate the ACE2 expression, revealing potential risk factors for severe illness from COVID-19. Here, we highlight three ACE2 modulating conditions. Data from GSE89714 show up-regulated expression of ACE2 in cardiomyopathy. Data from GSE104177 showed that RAD140, a selective androgen receptor modulator, induces ACE2 expression in human breast cancer xenografts. Data from GSE114013 show that Itraconazole, an antifungal drug, up-regulates ACE2 expression in two colorectal cancer cell lines, HT55 and SW948. These findings suggest that these drugs should be studied with respect to ACE2 expression in lung and cardiac cells and tissues and that patients on these drugs should be studied closely during the pandemic. Our GENEVA analysis is consistent with an increased death rate in COVID-19 patients with heart conditions[2–4] and suggests that higher ACE2 expression can contribute to the increased risk. We identified 3936 COVID19 patients from the electronic health records (EHR) of the University of California San Francisco (UCSF) hospital. We first compared the COVID-19 patients with cardiomyopathy (N = 43) to patients without cardiovascular diseases (N = 3269) and confirmed that cardiomyopathy is significantly associated with the risk of death (p = 0.004). We next compared the cardiomyopathy patients to patients with other cardiovascular diseases (N = 624). Multivariable Cox proportional-hazards regression confirms that cardiomyopathy is significantly associated with the risk of death (p = 0.038, 438 % increase in observed death rate).

References:
Image-based Phenotyping for Risk Stratification in the National Lung Screening Trial

Mirza S. Khan, MD1,2,3, Kaiwen Xu, MS2, Stephen A. Deppen, PhD3, Kim L. Sandler, MD3, Bennett A. Landman, PhD2,3
1US Dept. of Veterans Affairs, Nashville, TN; 2Vanderbilt University, Nashville, TN; 3Vanderbilt University Medical Center, Nashville, TN

Introduction
Annual lung cancer screening (LCS) with low-dose computed tomography (LDCT) is recommended for those at high risk for lung cancer based on age and smoking history. Yet, there remains an increased incidence in lung cancer among groups that fail to meet existing LCS criteria. We use image-based clustering analysis of complete thoracic images from high risk patients to identify sub-phenotypes at highest risk for lung cancer.

Methods
We randomly selected 1,200 male subjects with biopsy-proven diagnoses as reported in the National Lung Screening Trial (NLST); 200 of whom had biopsy-confirmed lung cancer. The first LDCT scan for each subject was pre-processed and then rigid registered to a common reference space as in Xu et al. We evaluated each rigid registered image and excluded 14 images due to missing lung fields or malorientation. We applied intensity-based hierarchical agglomerative clustering with complete linkage and mean absolute difference as our pairwise distance measure as described by Jin et al. We identified the optimal number of clusters with the elbow location of residual dissimilarity. We analyzed patient characteristic differences across clusters using the Kruskall-Wallis test for continuous variables and Chi-squared test for categorical variables. We studied the time to cancer diagnosis for each cluster using the Kaplan-Meier (KM) method and applied the log-rank test to assess for inter-cluster survival difference. To study the effect of each cluster on incident lung cancer diagnosis events, we used the Cox proportional hazards (CPH) model.

Results
Median body mass index (BMI) was lowest among those in Cluster 4 and Cluster 2 at 22.4 and 25.8, respectively. Lung cancer diagnosis rates differed among the clusters (p = 0.04) and was highest among Clusters 2 and 4; 20% in both. KM curves shown in Fig. 1A; log-rank test: $\chi^2 = 10.23$ (p = 0.04). As shown in Fig. 1B, Cluster 2 has a hazard ratio (HR) of 1.61 (95% CI 1.08, 2.38, p = 0.02) and Cluster 4 HR = 1.67 (0.98, 2.86; p =0.06).

Figure 1: (A) Kaplan-Meier curves for each cluster. (B) Hazard ratios from Cox Proportional Hazards model.

Conclusion
In this preliminary work, we identified 5 clusters, 2 of which demonstrate reduced time to lung cancer diagnosis. We find that these high-risk phenotypes may manifest clinically relevant features known to be associated with increased lung cancer mortality, e.g. low BMI. This work may be useful to guide consideration of other risk factors for LCS.

References
Dissemination and Implementation of a Child Abuse Clinical Decision Support System

Sundas Khan, MD1,2, Isabel Barata, MD2, Dana Kaplan, MD2, Safiya Richardson, MD, MPH1,2, Francesca Bullaro, MD2, Sera Levy, BA1, Fatima Malik, MHA1, Thomas McGinn, MD, MPH3, Emily Heineman, MA4, Rachel P. Berger, MD4

1Center for Health Innovations and Outcomes Research, Feinstein Institutes for Medical Research Hofstra/Northwell, Manhasset, NY; 2Donald and Barbara Zucker School of Medicine at Hofstra/Northwell, Hempstead, NY; 3iQuEST, Baylor College of Medicine, Houston, TX; 4UPMC Children’s Hospital of Pittsburgh, Pittsburgh, PA

Description of the Problem
Child abuse is an important cause of morbidity and mortality; evidence-based recommendations for screening children with suspected abuse is important for early recognition. The objective of the study was to disseminate an electronic health record (EHR)-embedded child abuse-clinical decision support system (CA-CDSS) developed in the Cerner EHR into the Allscripts EHR.

Methods
The CA-CDSS was disseminated into Northwell Health’s EHR at 4 emergency departments (ED) including a tertiary care pediatric hospital, 2 tertiary care adult hospitals, and a community-based hospital. After meetings with key stakeholders at each ED, a prototype was developed and refined based on usability testing with end users in less than 6 months. The CA-CDSS consisted of: 1) trigger system (including high-risk chief complaint triggers and a 5 question child abuse screen (CAS) (questions below) for children less than the age of 13 years), 2) ED status board alert, 3) provider alerts within clinical documentation, and 4) injury-specific order sets based on American Academy of Pediatrics guidelines. Use of the screen and positive screens was collected from June 2019-April 2020 (live period).

CAS Questions
1) For children presenting for evaluation of a possible injury, was there a possible or definite delay in seeking medical attention given the severity of the injury/injuries?
2) Are you concerned that the history may not be consistent with injury or illness?
3) Is the child developmentally “cruising” or walking? (different question(s) appear(s) based on response selection of yes or no)
4) Are there findings that might reflect poor supervision, care, nourishment or hygiene?
5) Are there any additional comments or concerns related to child abuse or neglect and/or additional explanations for any “yes” responses above?

Results
During the live period, the CAS was completed during 71% (33,876/47,943) of all patient visits of children < 13 years old; a Completion of the CAS was 64% in June 2019 and 84% in April 2020. The percentage of CAS that was positive was 1.6% (552/33,876).

Conclusions
The increase of the CAS completion rate at the end of the live period was sustained due to continuous monitoring, refresher trainings, and utilizing the CA-CDSS as a communication tool between nurses and providers. The CAS was completed at a high rate by ED nurses to aid in identification of children who should be further evaluated by ED providers for possible abuse and neglect. Each ED had a different clinical setting and it was important to evaluate site specific post-implementation strategies. Prior work in this area has focused on integration of guidelines into the workflow as well as order sets in the EHR. Our work includes utilization of a “trigger” system in the EHR to identify children for further assessment by selection of high-risk chief complaints and/or a positive CAS. Additional data for the CAS affecting rates of referral to child protective services during the baseline and live period is forthcoming.

Funding
Research reported in this work was funded through a Patient-Centered Outcomes Research Institute (PCORI) Award (DI-2017C1-6215). The views, statements, and/or opinions in this work are solely the responsibility of the authors and do not necessarily represent the views of PCORI, its Board of Governors or Methodology Committee.

References
COVID-19 screening system utilizing daily symptom attestation helps identify hospital employees who should be tested to protect patients and co-workers

Ellen Kim, MD, MPH1, Charles Morris, MD, MPH1, Michael Klompas, MD, MPH1, Haipeng Zhang, DO, MMSc1, Adam Landman, MD, MS, MIS, MHS1, Sunil Eappen, MD, MBA1, Hojjat Salmasian, MD, MPH, PhD1

1Brigham and Women’s Hospital, Boston, MA

Introduction
During the first few months of the SARS-CoV-2 (COVID-19) pandemic in 2020, there was uncertainty about the rate of COVID-19 transmission, transmission precautions needed for suspected and confirmed COVID-19 infections, and supply and accuracy of diagnostic tests. Consequently, symptom screening was used as a key infection control mechanism. The effectiveness of a symptom-based screening of healthcare workers for COVID-19 was unknown but required by a Massachusetts Department of Public Health mandate.

As previously described by Zhang et al1, our institution rapidly deployed a novel employee screening system that included a digital daily symptom attestation application called COVID Pass, whose source code was also shared online so other institutions could use it. All employees were screened with COVID Pass before entering the hospital or clinic for the day. Employees who attested to having symptoms were evaluated by our Occupational Health and Safety (OHS) team before being cleared for work or undergoing additional testing. Employees who were concerned about a potential infection were also encouraged to skip the attestation process and contact OHS directly.

Methods
We performed a retrospective cohort study of all employee attestations and COVID-19 tests at our institution. Descriptive statistics (chi-squared, Wilcoxon, and Kruskal-Wallis tests) compared characteristics of employees who submitted symptomatic attestations and tested positive for COVID-19.

Results
After data linkage and cleaning, we were able to study 2,117,298 attestations submitted by 65,422 employees between March and June 2020. Most attestations were asymptomatic (99.9%). Among the 2,026 symptomatic employees, the most common symptom was sore throat (25%); the most common symptoms associated with a positive COVID-19 test were anosmia (23% vs 4%) and fever (46% vs 19%).

We identified 905 employees who were tested within 14 days of a symptomatic attestation, and 13% of these tested positive. This is significantly higher than the <0.5% positive rate in the community and multiple healthcare setting scenarios2 and slightly higher than the reported 10% positive tests among symptomatic health care personnel (HCP) reported by one of the small departments at our institution during a much shorter period of time3.

Conclusion
Our daily symptom attestation system seems to help identify employees who should be tested for COVID-19 while preserving scarce tests and reducing exposure to patients and co-workers. It may also help unwell HCP to stay home from work, and could even help to start changing the prevalent culture of presenteeism in healthcare. Our results also support the continued use of daily symptom screening of patients and visitors before entering a hospital/clinic, to better protect themselves, patients, and HCP.

References
Information Needs Analysis of Diabetes Self-Management Education and Support for Older People with Diabetes

Min Soon Kim, PhD1,2, Ploypun Narindrarangkura, MS, MD1, Suzanne A. Boren, MHA, PhD1,2, Uzma Khan, MD3, Eduardo J. Simoes, MD, MSc, MPH1,2,
1Department of Health Management and Informatics; 2Institute for Data Science and Informatics; 3Department of Medicine, University of Missouri, Columbia, MO, USA

Introduction

Over one-fourth of Americans aged 65 years or older have diabetes. Diabetes self-management education and support (DSMES) has become a significant component of diabetes care for older people with diabetes. The AADE7® Self-Care Behaviors® (AADE7®) is a patient centered DSMES guideline for people with diabetes. AADE7® has seven principles: Healthy Eating, Being Active, Monitoring, Taking Medication, Problem Solving, Reducing Risks, and Healthy Coping. Successful DSMES can improve glycemic control and health outcomes. However, it is unknown whether diabetes providers are aware of DSMES and specifically AADE7® principles for providing care. In this study, we conducted a survey among providers with experience of providing care to adults with diabetes aged 65 and over to determine the acknowledgment and gap of AADE7® and DSMES. The long term goal of this study is to develop a patient centered DSMES intervention integrated into the clinic note to improve care of people with diabetes.

Methods

We invited family and community medicine providers and diabetes care specialists at the University of Missouri Health Care. An electronic survey was sent through the Research Electronic Data Capture (REDCap). The survey consisted of two parts: demographic and evaluation of randomly selected deidentified representative clinic note of older patient with diabetes. We collected information on age, gender, race, work experience, acknowledgment of AADE7® and DSMES. The clinic note evaluation part consisted of four questions: Q1) Does this note address the AADE7® principles? Q2) On a Likert Scales of 1 to 5, please mark if the note provides helpful information for patients about each AADE7® principle. Q3) Do you think the information in this note would be easily understandable by a patient with diabetes? Why? Q4) Do you have any additional comments on the information that may make this note address AADE7® principles to improve diabetes care?

Results

We analyzed responses of 42 providers (response rate 43%). Mean age was 52 (SD 14). 22 providers (52%) were female, and the average diabetes care experience was 23 years (SD 12). 25 providers (60%) were familiar with general DSMES guidelines; however, 27 providers (64%) were not familiar with AADE7®, 31 of 42 providers completed the clinic note evaluation. Per Q1, most providers (28, 90%) agreed the clinic note addressed the AADE7 principle of Taking Medication, but not the Healthy Coping principle (3, 10%). Per Q2, the parts of the note providing helpful information were Healthy Eating (2.2±1.0), Being Active (2.3±1.0), Monitoring (1.8±0.9), Taking Medication (2.9±1.1), Problem Solving (1.5±0.9), Reducing Risks (2.0±1.2), and Healthy Coping (1.5±0.9). Per Q3, 77% of providers agreed that the information in the clinic note would not be easy to understand for people with diabetes. 17 of 30 providers commented the excessive use of “abbreviations” would be the main reason prohibiting patients from understanding the clinic notes. Per Q4, 5 of 17 providers proposed a note template that outlines the seven principles in the note and would remind them to discuss each principle with patient at each visit.

Conclusion

This study investigated the acknowledgment and gap of AADE7® during clinic visits of older people with diabetes. Our preliminary results by providers showed that the current clinic note did not address important DSMES or AADE7® principles. Providers had difficulty delivering DSMES and AADE7® to older people with diabetes via a clinic note. We suggest that providers should consider readability for seniors and provide patient centered DSMES intervention at a fourth- to sixth-grade reading level, the following NIH guidelines. We are currently conducting a survey of older people with diabetes to determine patients’ information needs on DSMES to be integrated into the clinic notes. The comprehensive information needs from both groups will inform development of patient-centered clinic notes for the older people with diabetes.

Christopher Kitchen, MS¹, Hsien-Yen Chang, PhD¹, Jonathan Weiner, DrPH¹, Hadi Kharrazi, MD PhD¹,²

¹Center for Population Health IT, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA; ²Division of Health Sciences and Informatics, Johns Hopkins School of Medicine, Baltimore, MD USA

Abstract

Patient vital signs are underutilized in health risk assessment using electronic health records. Using an added value framework, we evaluated patient body mass index (BMI) and blood pressures (BP) and found improvement in predictions of patient cost and utilization over a commonly used tool, the Charlson comorbidity index. Healthcare systems and physicians could benefit from incorporating patient vital signs into their stratification of patient risk, especially in situations where little other diagnostic or historical information is recorded.

Introduction

Electronic health records have been integral in advancing health risk assessment and modeling, but health systems have not always made use of the very granular information stored in them.¹² This may especially be true in the case of patient vitals and laboratory findings. We believe these data are increasingly valuable in informing clinical decision support. We sought to replicate prior findings with respect to the added predictive value of body mass index (BMI) and blood pressure (BP) markers in health risk and utilization modeling. We also explored alternative definitions reflecting change in vital signs over time using this added value framework.

Methods

This study relied on a retrospective cohort design using electronic health records (EHR) records from Johns Hopkins Community Physicians (JHCP). Data reflect multiple emergency and outpatient care facilities throughout the Baltimore-Washington MSA. Structured patient level annual summaries for 2016 and 2017 were developed using the Johns Hopkins ACG System (v11.0). Patient Charlson comorbidity index and scores were attached to the patient summary file using the Washington MSA. Structured patient level annual summaries for 2016 and 2017 were developed using the Johns Hopkins Community Physicians (JHCP). Data reflect multiple emergency and outpatient care facilities throughout the Baltimore-Washington MSA. Structured patient level annual summaries for 2016 and 2017 were developed using the Johns Hopkins ACG System (v11.0). Patient Charlson comorbidity index and scores were attached to the patient summary file using the comorbidity package of the R programming language. 12,820 total patients met inclusion criteria and had at least one recorded value for BMI and BP in 2016. Values for patient height, weight, BMI, systolic and diastolic blood pressures were extracted from EHR flowsheet tables. Raw values were then transformed to a common scale for each measure and aggregated across patients by year of observation. Annual means for BMI and BPs were used to classify patient obesity and hypertension risk. Change (i.e. annual variability) was defined as the difference between annual maximum and minimum values for BMI and mean arterial pressure (MAP). Added value was assessed through model performance estimates using repeated 10-fold cross validation compared to base model point estimates for: (1) patient age & sex, (2) age, sex and the Charlson weighted index, (3) age, sex and the ACG system’s Dx-PM risk score. Added value of categories and change were evaluated as percent improvement in R² and AUC over base in predicting key outcome metrics, including total, medical or pharmacy costs and whether patients had concurrent or prospective ED visits, inpatient hospitalizations, and fell within the top 5% of sample for total costs.

Results

The largest subgroup of either classification is characterized as overweight (26.7%) and having elevated BP (46.8%). Both categorical BMI and BP were progressively indicative of disease comorbidity, but not uniformly related to healthcare utilization or cost. Categorical BMI and BP only significantly improved R² or AUC performance metrics compared to base models including just age & sex. Annual change BMI and MAP independently and additively improved concurrent year predictions over the Charlson weighted index for annual medical cost, emergency room visits, hospitalization and whether a patient was among the top 5% of patients by total annual cost.

Conclusion

Healthcare systems can leverage patient vital signs to improve risk assessment models but may experience diminishing returns when other factors are considered. These measures have little added value for models including commonly used indices of comorbidity and risk, though observed change or volatility in vitals over time may still yield positive results over diagnosis-derived risk scores alone. When a healthcare delivery system does not have other risk assessment information for a patient, recorded vitals may allow some simple estimation of disease risk, cost and certain outcomes.

References


Combining Chart Review and Hospital System Dynamics for Electronic Health Record Phenotyping in an International COVID-19 Research Network

Jeffrey G. Klann, PhD<sup>1</sup>; Griffin M. Weber, MD, PhD<sup>1</sup>; Emma Perez, MGC<sup>1</sup>; William Yuan, PhD<sup>1</sup>; Gabriel A. Brat, MD<sup>2</sup>; Shawn N. Murphy, MD, PhD<sup>3</sup>

<sup>1</sup>Harvard Medical School; <sup>2</sup>Research Information Science and Computing, Partners Healthcare; <sup>3</sup>Laboratory of Computer Science, Massachusetts General Hospital; <sup>4</sup>Beth Israel Deaconess Medical Center; <sup>5</sup>Brigham and Women’s Hospital–All in Boston, MA

Introduction

The coronavirus disease 2019 (COVID-19) pandemic has highlighted both the promise and current challenges in electronic health record (EHR) research. The Consortium for Clinical Characterization of COVID-19 by EHR (4CE) is a recently-convened international consortium of over 350 hospitals across nine countries on four continents.

Here, our goal is to calibrate our analyses by developing a multi-site proxy measure to detect whether COVID-positive patients are hospitalized for a COVID-related pathology. As hospital systems begin to open up, more patients are being admitted for routine surgeries. Because a COVID test is routinely performed at hospital admission, some patients will be COVID-positive asymptomatically during an otherwise routine hospitalization.

We apply a combination of chart review by clinicians at 4CE sites with an approach called “hospital dynamics.” In this approach, the analysis focuses on the metadata about a hospitalization rather than the treatments themselves. In previous work, we have found that meta-measures such as total number of lab tests on the day of admission, or time of day of lab tests can be highly predictive of disease course.

Methods

We defined a protocol in which clinical experts manually review the charts of COVID+ patients at their institutions and record a flag as to whether they were admitted for COVID or non-COVID reasons. We then utilize this manually annotated gold-standard to develop phenotyping algorithms for automatically predicting which hospitalizations are COVID-related.

We defined eligible patients for this study as those in the 4CE “COVID-19 cohort”: all hospitalized patients (both pediatric and adult) with their first positive test for SARS-CoV-2 +14/-7 days around the hospitalization.

Protocol details were based on clinical expertise and manually reviewed COVID+ patient charts at two hospitals. Our review found some complex cases in which it is hard to tell cases if a hospitalization is COVID-related, because COVID diffusely affects so many organ systems. Therefore, we chose a small set of criteria, instructing reviewers to label a hospitalization as COVID-related only if it included treatment or close monitoring for respiratory insufficiency, hemodynamic changes, or blood clots requiring significant anticoagulation.

We performed a preliminary hospital system dynamics study at one site, reviewing 59 charts from the COVID cohort during the first wave of the pandemic (March 13 to Jul 7, 2020) and recorded whether the hospitalizations were directly COVID-related, according to a review of the admit note and discharge summary. We examined a “Worry level” – the number of labs ordered for the routine vs. COVID hospitalizations.

We are presently in the process of performing a multi-site extension of this work across the full time-period of COVID. Sites will choose a random sample of 20 patients per calendar quarter and record a flag as to whether they were admitted for COVID or admitted for routine reasons. Then we will use hospital dynamics to finalize a proxy measure of routine care, including methods like frequency, association testing, and regression analysis.

Results

Our preliminary hospital dynamics study of 59 patients found that 18% of hospitalizations among COVID-positive patients were unrelated to COVID. We found that the total number of labs ordered on the first day of hospitalization was higher in the COVID hospitalizations. Additionally, when filtered to labs that we have found most correlated with severe COVID, there was an even larger difference. 94% of patients with a COVID-related hospitalization had been ordered LDH, procalcitonin, PaO2, PaCO2 within +/-2 days around hospitalization, whereas only 36% for the non-COVID related hospitalizations. We are in the process of reviewing charts at several more sites and will present further results in November at the Symposium.

Discussion

As hospitals return to normal care patterns, a COVID-positive test does not necessarily mean that COVID affected the patient’s hospital course. Here we employ chart review and hospital system dynamics techniques on an international federated COVID-research network.
U.S. COVID-19 Surveillance in PCORnet®

Sheryl A. Kluberg, PhD¹, Thomas W. Carton, MS, PhD², L. Charles Bailey, MD, PhD³, Julia A. Fearrington, MPH⁴, Keith A. Marsolo, PhD⁴, Kshema M. Nagavedu, MPH¹, Jon Puro, MPA, HA⁵, Jason P. Block, MD, MPH¹

¹Harvard Medical School and Harvard Pilgrim Health Care Institute, Boston, MA; ²Louisiana Public Health Institute, New Orleans, LA; ³Children’s Hospital of Philadelphia, Philadelphia, Pennsylvania; ⁴Duke University, Durham, North Carolina; ⁵OCHIN, Inc., Portland, OR

Introduction

The COVID-19 pandemic prompted the need for wide-reaching national surveillance that could rapidly adapt to accommodate new knowledge, policies, therapeutics, and vaccines. PCORnet®, the National Patient-Centered Clinical Research Network, a distributed data network funded by the Patient-Centered Outcomes Research Institute (PCORI), responded to this need with a focused update of highly current EHR data beginning in April 2020.

Background

PCORnet® is composed of 70 institutions (DataMarts) that streamline their data from diverse EHR sources into the PCORnet Common Data Model, typically refreshed quarterly. PCORnet was uniquely suited to rapidly establish a COVID-19 surveillance system, with infrastructure in place to pull and transform data and flexibility to examine new data elements. Governance for sharing aggregate findings was simplified by PCORnet’s distributed framework, with data maintained behind each institution’s firewall. PCORnet also has existing modular analytic tools that perform across DataMarts’ varied computing environments.

Methods

Forty-two EHR-based DataMarts contributed to the PCORnet® COVID-19 surveillance system, which was funded by the Centers for Disease Control (CDC) beginning in October 2020. DataMarts updated their data biweekly with frequent queries to assess characteristics of patients with COVID-19. To avoid an overload of data, only patients with respiratory illness ICD-10-CM codes (including COVID-19) or lab orders for SARS-CoV-2 tests were included.

We ran seven queries against the PCORnet® COVID-19 dataset between October 2020 and February 2021. Each query separately assessed children (<20 years) and adults, according to COVID-19 diagnoses, laboratory results, and care setting. We stratified these cohorts by age category, sex, race, ethnicity, month, geography (state), weight status, and smoking status. We also assessed comorbidity history and COVID-19-related symptoms and treatments for all cohorts.

Results

As of February 2021, we identified 820,801 adults and 162,471 children with a COVID-19 diagnosis code or positive laboratory result. The PCORnet® case counts over time and by race reflect national trends reported by the CDC. Patients of Black or African American race comprised 15% of all COVID-19 diagnoses, but represented 24% of inpatient diagnoses and 26% of patients treated with mechanical ventilation. During periods of greater COVID-19 prevalence, inpatient and emergency department diagnoses were more common. We observed trends in therapeutic use over time; for example, remdesivir and novel therapeutic use increased, while hydroxychloroquine use decreased.

Conclusion

This work has demonstrated that PCORnet® was well-suited to rapidly initiate COVID-19 surveillance, and that the network’s large patient population can provide timely and customizable insights into national trends over time.

References


Understanding Enterprise Data Warehouses to Support Clinical and Translational Research: Initial Findings on Enterprise Information Technology Relationships, Data Governance, Workforce, and Cloud Computing

Boyd M. Knosp, MS¹, Catherine K. Craven, PhD, MA, MLS², David A. Dorr, MD, MS³,⁴, Elmer V. Bernstam, MD, MSE, MS⁵, Thomas R. Campion, Jr., PhD⁶

¹Roy J. and Lucille A. Carver College of Medicine and the Institute for Clinical & Translational Science, University of Iowa, Iowa City, IA; ²Department of Population Health Sciences, University of Texas Health Science Center at San Antonio, San Antonio, TX; ³Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University, Portland, OR; ⁴Department of Medicine, Oregon Health & Science University, Portland, OR; ⁵Center for Clinical and Translational Sciences, University of Texas Health Science Center, Houston, TX; ⁶Clinical & Translational Science Center, Weill Cornell Medicine, New York, NY

Introduction

In a prior study, we illustrated how Clinical and Translational Science Award (CTSA) hubs have implemented enterprise data warehouses for research (EDW4Rs) with respect to organizational and technical architecture, processes for access, and service management [1]. In this study, we addressed gaps in understanding EDW4R operations regarding enterprise information technology (IT) relationships, data governance, workforce, and cloud computing.

Methods

We conducted semi-structured interviews via Zoom with a convenience sample of EDW4R leaders from 20 CTSA hubs. Data collection occurred between January and October 2020, and the study team conducted a directed content analysis of interview transcripts.

Results and Discussion

For enterprise IT, CTSA hubs collectively referred to infrastructure services provided by health system, university, and, in some cases, medical school IT organizations. 75% (n=15) of respondents stated that the team providing EDW4R service at their institution was in an organizational unit, such as a CTSA-funded research institute, separate from an enterprise IT organization. Some respondents (n=3) described enterprise IT as considering EDW4R needs as “not a priority” or a “stepchild” to health system needs. For data governance, most hubs (n=17) indicated some level of existing data governance to make decisions regarding EDW4R operations. More than half (n=12) indicated there were multiple levels of data governance, such as an oversight group to review external data sharing agreements and an operationally-focused team to review internal requests for data access. To inform how to add data to an EDW4R, eight hubs described user requests as dictating priorities while two indicated that the Observational Medical Outcomes Partnership (OMOP) common data model provided a guide. Regarding workforce, multiple roles comprised EDW4R teams, including managers who provided oversight and budget control; engineers who delivered data pipelines, extract-transform-load (ETL) of data, and technical operations; analysts who engaged with faculty, managed project deliverables, and provided phenotyping services; and faculty who served as clinical or methodological (e.g., machine learning) domain experts. Respondents valued staff versatility with one participant stating “we always need ninjas” and another describing the ideal hire as “more of a Swiss army knife than a dedicated developer.” For cloud computing, implementation for EDW4R has not been widely adopted with only two hubs describing cloud-based EDW4Rs. Some hubs reported strategic plans to move their EDW4R into the cloud, typically with the intent to restructure it into a more cloud-native architecture such as a data lake. Several hubs described using cloud for data enclaves and software as a service for CTSA-specific tools. Using qualitative methods this study characterized trends and opportunities for improvement that may contribute toward best practices for EDW4R operations and inform hubs.

Acknowledgements

Support for this study included UL1TR002537, UL1TR000457, UL1TR002369, UL1TR003167 and UL1TR001433.
Building with CEDAR and Making Evidence More FAIR

Peter Krautscheid, BS1, Kathy Mikk, JD, MFA1, Mario Teran, MD2, Edwin A. Lomotan, MD2

1The MITRE Corporation, McLean, VA; 2Agency for Healthcare Research and Quality, Rockville, MD

The problem

Considerable delay exists between the emergence of new research evidence, such as patient-centered outcomes research findings, and incorporation of the evidence into routine clinical practice. One reason for this delay is that informaticists and others often rely on time-consuming, manual processes to discover and retrieve the evidence, which must be further transformed into useful, actionable recommendations (e.g., via clinical decision support [CDS] tools).

The project

The Agency for Healthcare Research and Quality (AHRQ) is the trusted home of several evidence repositories that contain systematic reviews, care recommendations, and CDS artifacts maintained by its Center for Evidence and Practice Improvement (CEPI). The CEPI Evidence Discovery And Retrieval (CEDAR) project aims to make evidence more FAIR (Findable, Accessible, Interoperable, and Reusable) by prototyping a reference implementation (RI) using a standards-based application programming interface (API). The API will be publicly-available for other platforms to discover and retrieve AHRQ evidence across multiple repositories simultaneously, making it easier and faster for others to use AHRQ evidence to inform healthcare decision-making.

The analysis

The first task under CEDAR was to understand the current FAIRness of healthcare evidence repositories by evaluating the technical infrastructure and content of the CEPI repositories and relevant standards. The analysis included literature and web reviews; interviews with CEPI repository stewards (1-3 key informants each); review of supporting documentation (e.g., API documentation); and review of FAIR data principles and 12 tools for assessing adherence.

Table 1. CEDAR Analysis and Findings

<table>
<thead>
<tr>
<th>CEDAR Activity</th>
<th>Findings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identify methods and tools to assess FAIRness</td>
<td>Existing tools are not healthcare-specific; need to develop FAIR assessment criteria better tailored to evaluate AHRQ CEPI repositories and CEDAR</td>
</tr>
<tr>
<td>Document data types in CEPI evidence repositories</td>
<td>Varied data exist at different stages of the research evidence life cycle, e.g., systematic reviews, clinical practice recommendations; clinical decision support (executable code)</td>
</tr>
<tr>
<td>Conduct technical analysis of CEPI repositories</td>
<td>Differences exist in terms of metadata, taxonomies (e.g., MESH), API and search capabilities, and content for download</td>
</tr>
<tr>
<td>Identify applicable health IT standards</td>
<td>Many are applicable, e.g., FHIR; FHIR Clinical Reasoning; Clinical Quality Language; SMART; Info Button; CDS Hooks; FHIR Clinical Guidelines; EBM-on-FHIR; others</td>
</tr>
</tbody>
</table>

Conclusions

Facilitators, challenges, and gaps exist to building a FHIR-based CEDAR RI. The primary technical challenges are building robust search functionality and maintaining reliable data feeds from API and non-API-supported repositories. Ongoing stakeholder input will help clarify needs and opportunities to improve FAIRness of AHRQ’s research evidence. Next steps include designing, testing, and piloting the CEDAR RI for specific use case(s). The CEDAR prototype will likely inform other, external efforts to mobilize computable biomedical knowledge and to build common knowledge platforms across disparate sources of evidence, particularly those already in the public domain.

References

Randomized user testing of recommender system clinical decision support

Andre Kumar MD MEd, Rachael C. Aikens, Jason Hom MD, Lisa Shieh MD PhD, Jonathan Chiang MPH, David Morales MS, Divya Saini MS, Mark A Musen MD PhD, Michael Baiocchi PhD, Russ Altman MD PhD, Mary K Goldstein MD MS, Steven Asch MD MPH, Jonathan H Chen MD PhD
Stanford University, Stanford, CA

Introduction - Modern clinical practice is compromised by undesirable variability due both to the intractability of manually assimilating vast bodies of information and consistently applying such knowledge at scale. Modern electronic health records (EHRs) create the opportunity for data-driven clinical decision support (CDS) that utilizes the collective expertise of many practitioners in a learning health system. Standard tools such as order set templates already reinforce consistency and compliance with best practices, but maintainability is limited in scale by a top-down, knowledge-based approach requiring the manual effort of human experts. We previously developed clinical order recommender algorithms to dynamically produce decision support content from the bottom-up by data-mining clinical data sources. Yet, open questions remain on how clinicians would react to such suggestions.

In this randomized user testing study, we sought to determine how machine learned clinical order recommender systems for electronic order entry for simulated inpatient cases would affect clinical users.

Materials and Methods - We recruited 43 licensed physicians with experience treating medical inpatients to use a clinical order entry interface for five simulated emergency medical cases, where the system dynamically suggests lists of clinical orders based on presenting diagnosis and orders placed thus far. Each physician-case was randomized whether to have access to our previously developed clinical order recommendation system. A panel of board-certified clinicians scored the primary outcome of whether the user’s orders were clinically appropriate on a consensus scale from -10 to +10 per item. Secondary outcomes included usage metrics and survey responses.

Results - Clinical Appropriateness scores for user clinical order decisions was higher with the recommender system on (incidence rate ratio (IRR) 1.06, 95% CI [1.01-1.12]), corresponding to a higher number of orders (IRR 1.09, 95% CI [1.01-1.17]), without a clear difference in mean score per order. Order suggestions from the recommender system were more likely to match physician needs than standard manual search, with fewer clicks to complete cases (IRR 0.90, 95% CI [0.83-0.99]). Physicians used the order recommender in 98% of cases possible, despite being completely optional. 95% of participants agreed the system would be useful.

Discussion - Healthcare systems can improve patient care through standardized tools like order sets, but these may not align well with individual cases. Guidance for up-to-date medical care must come from multiple sources, with a recommender system approach being able to automatically adapt to individual patient cases and emerging practices.

Our expert panel found physicians were placed more useful orders and no more likely to place harmful orders with the recommender. All physicians used the recommender options at least once, even though it was completely optional. In comparison, the physicians did not use manually authored institutional order sets in the vast majority of cases. The tool was positively received by participants, who identified clear benefits toward workflow.

This study is an important step towards automated systems anticipating clinical needs without users even having to ask. People may then feel like the computers are working for them, instead of the other way around.

Conclusions - A clinical order recommender system interface for common clinical scenarios seen in hospital medicine and emergency medicine is usable and accepted by physicians, without adversely affecting the quality of patient care decisions reflected in adverse or irrelevant orders.

References
Tracking the COVID-19 outbreak in India via Twitter

Sahithi Lakamana, MS¹, Abeed Sarker, PhD¹

¹Department of Biomedical Informatics, School of Medicine, Emory University, Atlanta, GA 30322, United States

Introduction

From March to late May 2021, a COVID-19 outbreak ravaged India resulting in the largest national lockdown in the world.¹ As of July, 2021, India reported over 31 million cases and over 400,000 deaths.² While traditional sources lagged in compiling aggregated data in India, Twitter provided near real-time insights about diverse aspects of the outbreak. Here we report our efforts to track the Indian outbreak from streaming Twitter data using natural language processing and machine learning methods. We have made the outputs available via a public, interactive dashboard.³

Methods

We collected data in real-time via the Twitter streaming application programming interface (API) using COVID-19-related keywords (e.g., COVID, COVID19, corona virus) and used the associated metadata to geolocate the sources. We performed analyses on several topics relevant to the pandemic. Outbreak timeline and location: We used the volume of COVID-19 related tweets over time geolocated from India to track the outbreak timeline. We compared the timeline with other events in India (e.g., opening of public places) and also tracked volumes at the state level when possible. Syndromic surveillance: We applied a previously-developed COVID-19 Twitter symptom lexicon⁴ to detect specific symptoms that were reported by users from India, using inexact matching to detect tweets that may contain misspelled symptoms. We also tracked references to black fungus, which surged during the outbreak. Emotion analysis: We performed linguistic emotion analysis of the tweets using the lexicon curated by the National Research Council, Canada, which contains terms related to anger, fear, anticipation, trust, surprise, sadness, joy, sentiment (negative and positive), and disgust.⁵ In addition to this, we also quantified the aggregated anxiety levels over time, as expressed by the tweets.

Results

Between February and June 2021, over 500,000 tweets about COVID-19 were geolocated to be from India, and over 3.5 million tweets were posted about the Indian outbreak from outside of India. The chatter about COVID-19 increased almost at the same time as the number of confirmed cases in India, with high correlation. In fact, a small bump in Twitter chatter was observed before confirmed cases started rising, perhaps indicating the data could be used to predict the upcoming outbreak. The top tweeting states were Maharashtra, Karnataka, Tamil Nadu, and Uttar Pradesh—states which also recorded the highest number of COVID-19 cases. There was also significant correlation in the state-level case numbers and the number of tweets emerging from those states. Fatigue, Dyspnea and Cough were the top reported symptoms emerging from India, and emotion analysis showed a surge in negative emotions in 2021 compared to the previous year. Anxiety levels and concerns about black fungus (mucormycosis) also surged—the former near the beginning of the outbreak (April) and the latter near the end (May). These increases coincided with the rise in COVID-19 case numbers and the disproportionate increases in black fungus infections later on in the outbreak, respectively.

Discussion and conclusion

Social media, specifically Twitter, chatter encapsulates an abundance of information regarding COVID-19. The knowledge contained within this resource can potentially be leveraged to obtain real-time insights about the current and future pandemics. While our methods and the results presented here focus solely on India, the same methods can be applied to conduct real-time surveillance in other countries, including the United States. Here, we have only presented a selected set of insights obtained from Twitter and further details can be found in our interactive public dashboard.³

References

COVID Deniers: Analyzing #Scamdemic and #Plandemic Tweets
Heather D. Lanier, MPH, Sameh N. Saleh, MD, Christoph U. Lehmann, MD, Richard J. Medford, MD
UT Southwestern Medical Center, Dallas, Texas

Background: Almost four billion people used social media in 2021 with the average user holding almost nine accounts on different platforms.1 The use of social media during the COVID-19 pandemic has led to an "infodemic" generating mis- and disinformation with potentially grave consequences. To explore means of countering misinformation, we analyzed tweets containing the hashtags #Scamdemic and #Plandemic.

Methods: On Jan 3, 2021, using a Twitter scraping tool called twint, we collected 419,269 English-language tweets that contained "#Scamdemic" or "#Plandemic" posted between Jan 1 and Dec 31, 2020. On Jan 15, 2021, we used the Twitter application programming interface (API) to extract the same tweets (by tweet ID) and additional relevant user metadata. We explored descriptive statistics of tweets including their content and user profiles and determined tweet availability in both datasets. We used natural language processing tools to automatically label tweet sentiment polarity and manually reviewed a subset of tweets to verify sentiment.

Results: In 2020, 60,316 accounts tweeted 310,327 times using our selected hashtags. Frequency of tweets peaked in May 2020 and decreased steadily thereafter. Accounts had a median of 241 followers (interquartile range [IQR], 62 - 901) and 534 accounts (0.9%) were verified. Overall, 22.4% (n = 69,659) of tweets had retweets and 43.1% (n = 133,749) of tweets had likes. Mean weekly sentiment was briefly positive for #Plandemic when the pandemic was declared in March 2020 but was pervasively negative throughout the rest of the study period (Figure 1). Comparing the tweets from the Twitter API to those originally extracted through twint, only 37.5% (n = 77,285) of the tweets remained accessible on Twitter for #Plandemic and 45.8% (n = 47,735) for #Scamdemic.

Discussion: While social media has democratized speech, it also permits users to disseminate potentially unverified or misleading information that endangers people’s lives and public health interventions. Characterizing tweets and users that use hashtags associated with COVID-19 pandemic denial is important for understanding the dispersion of misinformation. Sentiment analysis that worked well for previous COVID-19-related topics2,3 showed only mildly negative mean weekly sentiment, but manual review revealed a pattern of mislabeled positive tweets that were sarcastic in nature. For example, the clearly (for a human reader) sarcastic tweet “This is the Greatest Show! Grab your popcorn! #Plandemic” was attributed a positive sentiment. The majority (59.7%) of original tweets for both hashtags were later no longer accessible on Twitter either due to tweet or account deletion or account suspension. Future work will elucidate those inaccessible tweets due to account suspension versus tweet deletion and exploration of content of tweets to separate misinformation from tweets pointing out and disputing misinformation. We also plan to conduct demographic analysis of users to understand the use of bots and identify profiles more likely to propagate misinformation.

Limitations: Our study was limited by using convenience sampling of two hashtags to explore COVID-19 misinformation as well as limitations of the Twitter API to provide a relevant subset of all tweets. Despite using a tool for sentiment analysis modeled on social media text, we note partially inaccurate sentiment results because our collection contained many sarcastic tweets.

Conclusion: Tweets with the hashtags #Plandemic and #Scamdemic in 2020 had a predominantly negative sentiment or otherwise, were often sarcastic. In conjunction with the preponderance of inaccessible original tweets, we believe that at the time of the tweet, posters were in denial of the COVID-19 pandemic and seeking to disperse related mis- or disinformation.

Figure 1. Tweet Sentiment by week for tweets with the hashtags #Plandemic and #Scamdemic and for all tweets. Link addresses, references to accounts (signified by @'username'), and the hashtags #Scamdemic and #Plandemic were removed. Only tweets with more than two words were analyzed.

A New Corpus for Clinical Findings in Radiology Reports

Wilson Lau¹, David Wayne¹, Spencer Lewis MD¹, Özlem Uzuner, PhD², Martin Gunn, MB ChB¹, Meliha Yetisgen, PhD¹

¹University of Washington, Seattle, WA, ²George Mason University, Fairfax, VA

Introduction

Radiology reports contain a diverse and rich set of information about clinical abnormalities reported by radiologists. Natural language processing (NLP) offers a way to access this embedded information, which is critical to improve clinical care and enable secondary use applications. In this study, we proposed an event-based schema to capture details of two types of clinical findings (lesions and medical problems) in radiology reports. Our gold standard corpus contains 300 CT radiology reports randomly sampled from the University of Washington (UW) clinical repository. In this abstract, we describe the schema and the annotated gold standard corpus as well as present baseline named entity recognition (NER) and relation extraction (RE) results.

Methods

In our annotation schema, each event is represented in a frame that is characterized with a trigger and a set of attributes (Figure 1). The trigger is the key phrase indicating the event (shown in red in Figure 1). Attributes are fine-grained details that describe the trigger. Two annotators annotated 500 CT reports by using the BRAT rapid annotation tool. The inter-rater agreement on event annotations in 30 notes was 0.83 F1-score. The final gold standard corpus included 2344 lesion and 8065 medical problem event annotations (presented in detail in Table 1). To automatically extract event frames, we defined trigger and attribute extraction as NER tasks and linking the extracted attributes to triggers as a RE task. We split the gold standard with 4:1 ratio into training, test sets. We experimented with two neural NER architectures: (1) Bi-LSTM-CRF (NeuroNER) and BERT². We trained a BERT RE classifier in a multi-task fashion to predict relations between the triggers and the attributes.

We compared the performance of different BERT models (BERT base², BERT clinical³, BERT rad⁴). We report results of the best performing BERT model, BERT rad, which we pre-trained on 3 million UW radiology reports. BERT rad model also captured the in-domain knowledge and outperformed NeuroNER in NER tasks (Table 1). Baseline RE results are promising. As future work, we will expand our gold standard annotations and improve the extraction performance on the baseline results.

References

3. Alsentzer E, et al. Publicly Available Clinical BERT Embeddings, the 2nd Clinical Natural Language Processing Workshop: ACL; 2019

| Table 1: Annotation statistics and performance results (F1-score) for NER and RE tasks. |
|-------------------|-----------------|-----------------|-----------------|-----------------|
| Trigger (T) Attribute | Value | total # Gold standard annotations | NER (F1) | RE (F1) |
| Description (T) | - | 2344 | 0.88 | 0.90 | - |
| | | | | | 0.81 |
| | | | | | 0.78 |
| | | | | | 0.80 |
| | | | | | 0.80 |
| | | | | | 0.90 |
| | | | | | 0.90 |
| | | | | | 0.92 |
| | | | | | 0.91 |
| | | | | | 0.69 |
Evolving the PCORnet data pipeline for National COVID Cohort Collaborative (N3C)

Adam M. Lee1, G. Marshall Clark1, Sofia Z. Dard1, Kellie M. Walters1
1 University of North Carolina at Chapel Hill, Chapel Hill, NC

Introduction

The North Carolina Translational and Clinical Sciences Institute was called to submit recurring extracts from our PCORnet Common Data Model (PCDM) to the National COVID Cohort Collaborative (N3C). We aimed to deliver data weekly; however, the existing data pipeline took more than two days. The request forced us to evolve the data pipeline (i.e., Epic Clarity to PCDM) from a half-week, hands-on endeavor to an over-the-weekend automated task.

Methods

The environment and data pipeline involves 3 phases 1) create PCDM tables from SQL files [transform] using the source SQL Server, 2) extract those PCDM tables 3) load into an Oracle database. The first phase was to evaluate the data pipeline components used to populate the PCDM and determine where time savings methods could be utilized.

Previous State: A Python script was used to independently build each PCORnet table on a shadow copy server of Epic’s Clarity analytical reporting database (SQL Server). The completion time for this process was about 51 hours. The actual sequential run time totaled 33.83 hours, but the process had to be paused for nightly database maintenance, adding to the overall time. Next, SAS was used to extract data from SQL Server, load it into an Oracle 19c database, and generate SAS datasets for each table. This process took about 4.5 hours. Another SAS program performed data quality checks and migrated the Oracle Schema from staging to production. All these steps occurred in sequence.

Current State: Efforts were made to make processing parallel where possible and find faster data transfer methods. Python was chosen as the platform language to optimize the process. The Python code was revised to allow parallel table builds. Dependencies (e.g., demographic, encounter) were generated first. Then all independent tables (e.g., diagnosis, procedures) were created in tandem. Instead of outputting into a native SQL Server table, output was directly written to disk. We replaced SAS with Oracle SQL*Loader as the tool to load data into Oracle and noted faster speeds. Python was used to execute all commands with SQL*Loader. Finally, we created the SAS data files from the Oracle tables rather than embedding this step in the extraction process.

Results

The process optimizations resulted in a 70.28% decrease in time for the ETL process of the PCDM data pipeline.

<table>
<thead>
<tr>
<th>Table 1. Execution Times for PCORnet CDM for N3C specific dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transformation</td>
</tr>
<tr>
<td>Previous State</td>
</tr>
<tr>
<td>Current State</td>
</tr>
</tbody>
</table>

Discussion

One significant barrier that drove this solution was the topology of the analytical reporting database. In the previous state, database writes were permitted but slow. A new topology utilizing linked servers further made writes complex, but database reads more optimal. By opting for directly writing CSV files of the PCDM tables, we could bundle the extraction and transformation phases. Running the tables in parallel was the biggest time saver and allowed us to avoid nightly downtimes altogether. Further improvements could be gained if either the source data could host the PCDM data or switch to incremental data processing.

Conclusion

Overall, we were able to redesign and improve the data pipeline timeframe by utilizing parallel processing and tools more native to the technologies driving the databases. Streamlining processes allowed the removal of downtime and effectively delivered N3C data payloads without intermission or delays.
Predictive Modeling of Healthcare Utilization Metrics Identifies Adult Patients at High Risk for Suicide Attempt in the Primary Care Setting

Katherine Lee, B.S.¹, Colin G. Walsh, M.D., M.A.¹,²
¹Vanderbilt University School of Medicine, Nashville TN; ²Vanderbilt University Medical Center, Departments of Biomedical Informatics, Medicine, and Psychiatry, Nashville TN

Introduction: Suicide is one of the leading causes of death. Current guidelines recommend the use of screening tools such as the PHQ-9. However, these tools require time, have variable quality, and display limited sensitivity. A potential solution to the limited clinical utility of isolated predictors lies in machine learning, though challenges of overfitting and low precision modeling remain. One clinically implemented model has validated performance with feasible numbers needed to screen (NNS) and Area Under the Receiver Operating Characteristic Curve (AUROC) of 0.82 for a 30-day prediction window to automate “first-pass” screening for suicidality. To reduce gaps where these models do not perform equally well in all settings, e.g., in primary care, an improved understanding of setting-specific risk factors is necessary. To this end, we present a preliminary study of healthcare utilization in the primary care setting incorporating novel and nuanced feature engineering to identify usage patterns that contribute to suicide risk.

Methods: This is a retrospective cohort study of adults with a historic primary care encounter at Vanderbilt University Medical Center (VUMC) and a computed risk score for future 30-day suicide attempt from June 2019 to January 2021. Data were derived from the Research Derivative, a database containing clinical electronic health record data at VUMC. One hundred and four unique providers as well as admitting source and discharge location of encounters for the prior 6 months and 5 years. Cases were defined as patients with a risk score for future 30-day suicide attempt in the highest validated risk quantiles (above 1% predicted risk, NNS ~368) as validated prior. Controls were defined as those with a prediction of less than 0.01. The data were split into a random hold-out of 75% for training and 25% for testing. Feature importance was measured using SHapley Additive exPlanations (SHAP). Logistic regression hyperparameter tuning was conducted via grid search. All preprocessing and analyses were conducted in Python.

Results: We identified 13,718 risk score calculations for unique encounters, including 4,195 cases and 9,523 controls. A random forest model (RF) with 1000 trees outperformed logistic regression (LR) on the testing set: AUROC 0.89 (0.88-0.91; 95% bootstrap CI) for RF and 0.78 (0.76-0.80) for LR, Area Under the Precision-Recall Curve 0.82 (0.80-0.84) for RF and 0.66 (0.63-0.69) for LR. The SHAP summary plot is shown in Figure 1. “High” risk was most significantly predicted by few care sites and primary care encounters and many discharges to inpatient psychiatry.

Conclusion: A predictive model determined solely by metrics of healthcare utilization predicted suicide risk as “high” or “low” risk. This study builds upon pre-existing models by elucidating specific healthcare usage patterns that impact suicide risk. Further, studying the primary care population facilitates translation of computed risk to actionable interventions for lowering risk, which are frequently and primarily conducted in the primary care setting. Future directions include incorporation of these additional utilization metrics into the primary risk prediction model, inclusion of external data from other hospital systems and state department, and consideration of other algorithmic methodologies including deep learning.

References

Assessing Clinical Staff Usability & Satisfaction Before and After an Electronic Health Records Implementation Using Health-ITUES

Rachel Lee, PhD, RN\(^1\), Sarah Collins Rossetti, PhD, RN\(^1,2\), Jonathan Elias, MD\(^2,3\), Amanda Muy, MPH\(^2\), Eugene Lucas, MD\(^2,3\), Jessica Schwartz, MPhil, BSN, RN\(^1\), Erika Abramson MD, MSc\(^4\), Jessica Ancker, PhD\(^5\), Susan Bostwick, MD, MBA\(^4\), Kenrick Cato, PhD, RN\(^1,2\)

\(^1\)Columbia University School of Nursing, NY, NY; \(^2\)Columbia University DMBI, NY, NY; \(^3\)New York-Presbyterian Hospital, NY, NY; \(^4\)Weil Cornell Medical Center, NY, NY; \(^5\)Vanderbilt University Medical Center, Nashville, TN

Introduction: The usability of electronic health records (EHRs) is a significant contributor to documentation burden and burnout rates in clinical staff\(^1\). The implementation of a new EHR provides an opportunity to study staffs’ changing perceptions about usability. This study’s primary objective was to survey clinical staff before and after the implementation of a new vendor EHR and compare the changes in EHR usability. The secondary objective was to stratify the survey results by respondents’ roles and settings to describe the effects of individual characteristics on usability of the new system.

Methods: Anonymous electronic surveys were sent out to patient-facing clinical staff at a large academic medical center before and after the implementation of a new vendor EHR system. The survey was customized from the Health-IT Usability Evaluation Scale (Health-ITUES)\(^2,3\). Respondent demographics, including clinical role, specialty, work setting, and other EHR use history were gathered. Questionnaires were considered complete when the amount of missing data was <10%. One-way and two-way analysis of variance tests and pairwise comparisons with Bonferroni correction within groups were conducted on responses based on demographic information. Independent sample t-tests were conducted to assess the differences in means between pre-and post-implementation.

Results: A summary of results are displayed in Table 1 and Table 2. Tables are shortened for submission.

Conclusion: There were significant differences in usability perceptions before and after the EHR implementation across all constructs and by roles and settings. The overall mean of QWL, PU, and UC increased, and the overall mean of PEU and CSSA decreased after the implementation. The significant decrease in PEU were specific to RNs and MNAs while the decrease in CSSA were only observed among OPs and MNAs. The differences in changes reflect the varying experiences with the EHR by roles and settings. This result will inform and guide our future EHR optimization efforts by focusing on certain usability constructs, roles, clinical workflows, and settings.

References

Extracting Social Isolation Information from Psychiatric Notes in the Electronic Health Records

Lauren A. Lepow, MD1,*, Braja G. Patra, PhD2,*, Isotta Landi, PhD1, Prakash Adekkanattu, PhD2, Mark Olfson, MD/MPH3, J. John Mann, MD3, Euijung Ryu, PhD4, Joanna Biernacka, PhD4, Girish Nadkarni, MD, PhD1,4, Myrna Weissman, PhD3, Priya Wickramaratne, PhD3, Benjamin Glicksberg, PhD1, Jyotishman Pathak, PhD2,*, Alexander Charney, MD/PhD1,***

1Ichan School of Medicine at Mount Sinai, New York, NY, USA; 2Weill Cornell Medicine, New York, NY, USA; 3Vagelos College of Physicians and Surgeons, Columbia University, New York, NY, USA; 4Department of Quantitative Health Sciences, Mayo Clinic, Rochester, MN, USA

Introduction: Poor social support is a risk factor for worse outcomes in psychiatric populations1. The electronic health record (EHR) is a valuable repository of longitudinal data of psychiatric outcomes; however few patients have structured data regarding the social determinants of health. Clinical notes-- in particular psychiatric notes-- often contain information about psychosocial factors and have begun to be mined by natural language processing (NLP) techniques. Here we aim to validate a rules-based system for identifying social isolation with note templates on a document level.

Data: The corpus from the Mount Sinai Data Warehouse EHR is drawn from psychiatric inpatient, emergency room, and hospital consult encounters which consists of 286,692 notes for 33,800 patients. An advantage of the dataset is the frequency of clinical note templates containing four relevant questions: (1) ‘lacks social support: yes/no,’ (2) ‘social isolation or lack of social support network: yes/no,’ (3) ‘social isolation or lack of social service: yes/no,’ (4) ‘social isolation: yes/no’.

Method: Regular expression rules were implemented in the entire corpus to identify the social isolation templates (“temp-regex”) and were collected with the response ‘yes’ or ‘no’. A template was found in 41,190 notes, which became the corpus for the methods below. The templates themselves were then removed from the notes. We then developed a regular expression-based NLP system (“lex-regex”) to extract notes containing social isolation terms using a social isolation lexicon consisting of 23 words derived from Zhu et al.2 (e.g., ‘loneliness’, ‘limited social support’, ‘socially isolated’). We searched each lexicon term in our EHR system and observed its context in 100 unique notes. These contexts were reviewed, and based on domain knowledge, regular expressions were added and modified to exclude irrelevant and negated examples. The lex-regex is then tested for its performance measures by being evaluated against the ‘yes’/’no’ answer of the template.

Results and Discussion: The social isolation template was found in 41,190 notes; 1331 psychiatric notes from 552 patients contained both a lexicon seed term and template. See Table 1 for confusion matrix. The precision was 0.64 with a negative predictive value 0.686, specificity of 0.98, and recall of only 0.064. This pattern resulted from the low occurrence of lexicon hits compared to template occurrence, and is currently informing updates to the lexicon. The approach is being implemented at another academic medical center to be validated with patient survey data regarding social support and isolation. This proof-of-concept leverages semi-structured ‘yes/no’ clinical note data to help validate a regular expression-based NLP system which can eventually be used to extract social determinant information from the EHR. Future psychiatry research can utilize social determinant data in the context of other clinical and genetic information associated with the EHR.

References:


*co-presenters. **joint senior-authors.
MethylDrift: A Software Package for Calculating Age-related Epigenetic Drift in Human Tissues Using DNA Methylation Data

McKenna C. Lewis\textsuperscript{1,2}, Niema Moshiri, PhD\textsuperscript{1}, Kit Curtius, PhD\textsuperscript{2}

\textsuperscript{1}Department of Computer Science & Engineering, \textsuperscript{2}Division of Biomedical Informatics, Department of Medicine, University of California, San Diego, La Jolla, CA, USA

Introduction

Biological aging can be measured from changes in DNA methylation, and machine learning approaches exist to predict the chronological age of a patient based on a small number of CpG sites.\textsuperscript{1} Widespread age-related drift of methylation levels across the genome also influences accelerated biological aging and carcinogenesis,\textsuperscript{2} but it is not typically characterized in studies of differential methylation across tissue types. Our aims were to create a novel informatics tool for computational biologists to identify age-related tissue-specific epigenetic drift in a given dataset, to compute rates of aging across patient samples, and to ultimately apply the pipeline to any tissue of interest given user input.

Methods

We created a Python package called ‘MethylDrift’ that accepts either pre-processed or raw data from Illumina 450K or EPIC methylation arrays, extracts methylation values (\textit{m-values}) at over 450K CpG sites along the genome, and performs linear regressions to quantify age-related drift across a set of patient samples. Assessment can be stratified by sex, race/ethnicity, tissue phenotype/location, or other variables. The package also performs ANCOVA to determine differential drift with age between normal and (pre)cancerous tissues and identifies drifting vs. static CpG islands/sites. The output includes text files with static and drifting CpGs, drift rates of aging for use in downstream analyses, and a Jupyter notebook that renders graphs in a collection of vector PDFs in corresponding subdirectories.

Results

As a case study, we applied MethylDrift to publicly available normal colorectal tissue samples (\textit{n} = 264) and premalignant adenoma samples (\textit{n} = 159). Amongst a list of 12,700 validated drifting CpG sites in normal colon provided as input,\textsuperscript{2} we found 128 normally hypomethylated CpG sites (\textit{m-value} < 0) that undergo positive, accelerated differential drift in precancerous adenomas (Figure 1). Pre-cancer drift rates (mean = 0.017 \textit{m-val}/year, sd = 0.0097) were similar to those found in 322 colorectal cancers in The Cancer Genome Atlas (mean = 0.015 \textit{m-val}/year, sd = 0.0045). This implies accelerated aging begins early in progression and may play a role in malignant transformation, thus warranting consideration in biomarker studies as an included variable beyond patient age. The tool is available at \url{https://github.com/mckennalewis11/MethylDrift}, which also creates a Markdown file for results.

![Figure 1](image1.png)

**Figure 1.** The MethylDrift pipeline takes input sample data, selects CpG sites undergoing age-related drift, plots the results, and computes drift rate distributions that can be used as tissue-specific aging rates in studies of cancer risk.

Conclusion

MethylDrift provides a fast, comprehensive bioinformatics pipeline for the quantification of epigenetic drift in ultra-large datasets, which can then be used to create tissue-specific clocks for inferring ages of precancerous lesions.\textsuperscript{1} The main significance of this work is understanding the impacts of epigenetic drift on normal aging and cancer progression for future use in clinical practice. Integration with other multi-omics and clinical data is part of ongoing work.

References

Contextual Predictors of Medication Persistence for Patients on Oral Hypoglycemic Drugs

Zhiguo Li, Ph.D.1, Chandramouli Maduri, M.S.1, Ching-Hua Chen, Ph.D.1
1Center for Computational Health, IBM Research, Yorktown Heights, NY, USA

Introduction Oral hypoglycemic drugs (OHDs) are commonly prescribed to people with type 2 diabetes as a first-line of therapy to help control blood glucose levels. It has been shown that 55-65% of patients do not persist with their prescribed OHD within the first year of receiving a new prescription1. Using U.S.-based commercial healthcare claims and encounters data contained in the IBM® MarketScan® Research Databases, we predict the time to first non-adherence event for 46,450 diabetic patients who were newly initiated on an OHD. We examined the relative importance of static (i.e., ‘trait’-like) patient attributes and dynamic (i.e., ‘state’-like or contextual) patient attributes for this prediction task.

Method The index date (ID) for a patient is defined as the date of the first prescription filled for any OHD, with no observed claims for OHDs in the 365 days prior to the ID. Our study window spans 2.5 years of patient history (i.e., 12 months prior to patient’s ID, and 18 months after the ID). We use monthly Proportion of Days Covered (PDC) as a measure of adherence to OHDs in a given month. A PDC value of less than 0.8 is assumed to be indicative of non-adherence. We generated 60 baseline and 66 dynamic features from our data set. The baseline features are summary statistics computed over the 12 months prior to the ID. The dynamic features are time-varying variables that are computed for each month after the ID. The baseline features capture static, or ‘trait’-like patient attributes, while the dynamic features capture ‘state’-like, or contextual, patient attributes. We also include past values of PDC as features for predicting future PDC values. For each of the \( t = 1, 2, \ldots, 18 \) months in our prediction window, we trained a Random Survival Forest model, \( \text{RSF}_t \), to predict the time to first non-adherence. For a given month, \( t \), \( \text{RSF}_t \) is used to predict the time to first non-adherence, using only data from months prior to, and including \( t \). We experimented with training RSF models that utilized only the baseline features, only the dynamic features, only past PDC, and combinations thereof. We use Harrell’s concordance index (C-Index) as a measure of prediction performance.

Results Our results are summarized in Figure 1. The left plot shows that all models perform comparably, with C-index values increasing with \( t \). Additionally, \( \text{RSF}_t \) generates predictions with reasonable C-Index (i.e., above 0.7) when \( t > 3 \). In the right plot, we observe that as \( t \) increases, past PDC values increase in importance, while baseline features decrease in importance. Dynamic features remain consistently important, suggesting that while past behavior helps predict future behavior, contextual factors are also relevant, and may provide insights for intervention (e.g., providing extra support after a recent hospitalization, providing subsidies when there is an increase in medication cost, or simplifying the medication regimen when medication complexity increases significantly).

![Figure 1](image.png)

Figure 1: The plot on the left shows C-Index by prediction month, while the plot on the right shows the contributions of different feature types to model predictions.

References

An Exploration of Reasons Behind Drug De-escalation and Discontinuation Events in Clinical Notes

Jennifer J. Liang, MD, Diwakar Mahajan, MS
IBM TJ Watson Research Center, Yorktown Heights, NY

Introduction
Understanding the reasons behind the cessation or de-escalation of drug therapy is critical to medication management, allowing providers to better avoid adverse reactions, direct future treatment options, and improve adherence. However, such information is often buried in clinical notes and difficult to find. To automate extraction of this information, prior works have broadly categorized reasons behind drug discontinuation events, and also identified a need for more refined categories of discontinuation reasons. In this work, we explore more fine-grained reasons behind drug de-escalation and discontinuation events using the Contextualized Medication Event Dataset (CMED), and leverage the context dimensions in CMED (Temporality, Certainty, Actor) to provide further insights into these events.

Methods
CMED is a dataset capturing contextual information – Action (Start, Stop, Increase, Decrease, OtherChange, UniqueDose), Temporality (Past, Present, Future), Certainty (Certain, Hypothetical, Conditional), and Actor (Physician, Patient) – for medication changes documented in clinical notes. For this work, we used the subset of CMED where Action is either Stop or Decrease, consisting of a total of 475 medication mentions over 206 notes. We annotated each Stop or Decrease event with its corresponding reason span when available (Figure 1). Each reason span was then assigned 1 of 9 types emergent from the dataset: adverse event (anticipated, perceived, or confirmed), lack of efficacy, no longer need (e.g. due to clinical improvement, diagnosis change), planned regimen, planned procedure, medication change, expired prescription, insurance lapse, and other (e.g. NPO, closed formulary). We then analyzed the reason type along the CMED context dimensions.

Results
Out of 475 Stop or Decrease events, 206 do not have a specific reason for the event documented. The distribution of reason types for the remaining 269 Stop or Decrease events with reason documented are presented in Table 1. Overall, adverse event was the most common type, accounting for 50.2% of documented reasons. On analysis of reason type by context dimensions, we observed a couple interesting insights. First, when analyzing along the Temporality dimension, while Past and Present events have adverse event as the most common reason type, when limiting our analysis to only Future events, the most common reason type is a planned procedure (31.3%), followed by planned regimen (27.1%). Switching to the Actor dimension, when limiting our analysis to only patient-initiated actions (Actor: Patient), the most common reason type is an expired prescription (42.9%), followed by perceived adverse effect (28.6%).

Conclusion
We present an exploration of reasons behind drug de-escalation and discontinuation events documented in clinical notes along CMED context dimensions (Temporality, Certainty, Actor). Our analysis shows the potential of leveraging context dimensions to better understand reasons behind medication changes. In future, we plan to expand annotation to drug intensification events (Action: Start, Increase) in CMED. We are considering releasing these annotations when CMED becomes publicly available to further support development of an automated system for reason extraction.

References
Cycle-Consistent Adversarial network with criterion for COVID-19 Chest X-ray Generation

Zhaohui Liang, PhD, Jimmy Xiangji Huang, PhD
York University, Toronto, Ontario, Canada

Introduction

Chest X-ray and computed tomography (CT) imaging together with real-time polymerase chain reaction (RT-PCR) test are the commonly used fast screening methods for coronavirus disease 2019 (COVID-19) during the pandemic from early 2020. The COVID-19 positive cases have special bilateral or unilateral mottling and ground-glass opacity patterns on the chest X-ray and CT images. These patterns can be detected by deep neural network (DNN) images classifiers and integrated to the compute-aided health systems. Many successful studies on DNN models either for COVID-19 chest X-ray or CT image detection have been published. However, recently study revealed the seemingly high-performance deep neural networks (DNNs) models for COVID-19 chest X-Ray image detection are vulnerable from network attacks. One constraint is that they are optimized by extremely imbalanced datasets where the COVID-19 images only occupy 5% to 6% of the total. Another drawback is the specific image patterns of medical images, which are different from general-purposed images such as those in the ImageNet dataset. When using transfer learning with DNN models trained by the ImageNet to fine tune a new model for the radiography images, the pretrained feature extractors usually cannot effectively capture the medical significant patterns but developing meaningless pattern combinations. The above weakness contributes to the vulnerability of the current DNN models.

Cycle GAN with Criterion

We develop an enhanced cycle-consistent adversarial network (Cycle GAN) to generate synthetic COVID-19 chest X-ray images from normal images. The conventional Cycle GAN uses two generators and two discriminators to learn the mapping between two distributions by optimizing with a complex objective and reaching a state of adversarial equilibrium. However, the state of the adversarial equilibrium is uncontrollable and usually restricts the generators to reach a better optimal point. To increase the oscillation momentum of the adversarial equilibrium, we added a pretrained classifier with the accuracy of 99.7%, as the criterion to determine whether the synthetic images belong to the correct class, and to inject extra loss once every 5 steps during the Cycle GAN optimization. The new loss function for the generators has four terms: the adversarial loss, the cycle consistency loss, the identity loss, and the classification loss:

$$\mathcal{L}_{total} = \mathcal{L}_{GAN}(G, D_Y, X, Y) + \mathcal{L}_{GAN}(F, D_X, Y, X) + \lambda \mathcal{L}_{cycle}(G, F) + \lambda \mathcal{L}_{iden}(G, F) + \kappa (\varphi \mathcal{L}_{class})$$

where $\lambda$ and $\varphi$ are the parameters to respectively adjust the importance of the different loss terms during the whole Cycle GAN architecture optimization. The classification loss injection is controlled by a function $\kappa$ to determine the frequency of classification loss injection. We set $\kappa$ to be once every five steps in this work because too frequent injection of classification loss will shift the adversarial equilibrium and reduce the fidelity of the synthetic images. The generators follow the U-NET architecture with skip connections to reduce the input feature size from 128-by-128 to 1 by 1 then to restore to 128 by 128. The discriminators follow the PatchGAN architecture with an output of 8-by-8-by-1 feature map to determine with the images are real or fake. The newly added criterion uses the sparse categorical cross entropy as the objective for adversarial optimization, and the leave squared loss is used for the rest terms.

Experiments and Results

We use an COVID-19 image dataset consists of 219 COVID-19 positive images, and 1,064 normal chest X-ray and 1,064 viral pneumonia images to respectively optimize the new Cycle GAN with criterion model and a conventional Cycle GAN model without criterion. The results shows both Cycle GAN models can generate the complete landmark of the chest X-ray images. The conventional Cycle GAN (left in Figure 1) seems to produce more plausible COVID-19 image. However, the pretrained criterion accurately classifies 99.69% of the synthetic images by the new model compared to only 77.30% by the conventional Cycle GAN.

Conclusion

This study demonstrates the new Cycle GAN with criterion can generate more accurate COVID-19 positive chest X-ray images compared to the original Cycle GAN model. The criterion provides a control to the GAN architecture to generate desirable medical X-ray images with complex discriminative pattern associate with medical expertise.

This work is supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada and the York Research Chair (YRC) program.
Proactive Documentation Design to Meet Needs Now and Later
Zachary C. Liao, M.D., M.P.H.¹, Michelle Stoffel, M.D., PhD.¹, Michael G. Leu, M.D., M.S., M.H.S.¹,² J.P. Giliberto, M.D.¹, Justin Birge, M.D.³, Laura Higdon, C.P.C, C.P.M.A.¹, Angad P. Singh, M.D.¹

¹University of Washington, Seattle, WA; ²Seattle Children’s Hospital, Seattle, WA; ³University of Nebraska, Omaha, NE

Introduction:
Clinical notes, in addition to documenting care, must satisfy complex financial and regulatory needs. Most providers navigate these demands by creating their own note templates. This results in considerable variability in content and unnecessary data (“note bloat”), making review of these notes inefficient for clinicians, compliance, Health Information Management (HIM), and coders alike. Health system-level interventions have demonstrated promise for reducing documentation burden and associated burnout [1]. We leveraged our health system’s inpatient electronic health record (EHR) vendor transition to prospectively design effective standard notes with the goal of improving provider satisfaction and decreasing documentation time.

Methods:
To establish note standards, we assembled a committee of clinical, technical, and administrative stakeholders. We analyzed existing note templates for five note types (History and Physical (H&P), Progress Note, Consult Note, Discharge Summary, and Brief Operative Note), identifying common components that could be modularized and used across multiple notes. Inclusion of Compliance and HIM in the process ensured that regulatory standards were met during the clinical design phase. We adapted Agile methodology for our build process, employing cycles of rapid prototyping, implementation, and feedback [2]. Physician informaticists completed most of the tool build, which streamlined interactions with our clinical subject matter experts (SMEs). We also identified evaluation metrics to be collected in the months following EHR implementation on March 27, 2021. Measures included total time in the chart and in documentation activities, derived from user login data and direct observations.

Results:
Within five months, we designed and built 133 note templates across the five note types. Notes utilized common modular elements wherever possible, simplifying both maintenance and Compliance/HIM review. For example, our generic H&P template contained 17 modular components, each of which was used across multiple templates. Our “History of Present Illness” modular component was used in 43 templates. All templates were approved by clinical SMEs (representing 31 medical specialties), Compliance, and HIM.

We prioritized functionality and versatility in our standard notes. Collapsible note sections were employed to reduce visual clutter, with only the Assessment and Plan sections expanded by default. We created short and long options for medical history and objective data (e.g., lab, imaging), allowing writers to adjust the level of detail within the standard notes. Wherever possible, we embedded chart navigation links to allow writers to review data in flow without importing superfluous content into the note itself. Specialty-specific refinements were also made; for example, the Newborn H&P was configured to auto-populate labor and delivery information.

Discussion:
Effective documentation tools are critical, as the majority of provider time in the EHR is split between documentation and chart review [3]. Our modular standard note approach leveraged EHR functionality to address both clinical and non-clinical needs, with the added benefit of simplifying future maintenance. However, this strategy requires significant upfront planning and coordination among clinicians and other stakeholders, and the change management needed for implementation may be more challenging in mature EHR settings. Based on analogous research [4], we anticipate that notes created using our standard templates will be more acceptable for downstream clinical, administrative, financial, and technical uses.

References:

1725
Integrating Gaming into Virtual Medication Reconciliation Education
Veena Lingam, MD, Victor Garcia, MD, Chao-Wei Tsai, MD, Jacob Wooldridge, MD, Rachel Wong, MD, MBA, MPH
Stony Brook University Hospital, Stony Brook, NY, USA

Background: The COVID-19 pandemic created the need to re-formulate educational curricula for the virtual setting. With less opportunity for hands-on experience, simulation of essential clinical tasks using games are a potentially scalable way to both teach and evaluate trainees1. Medication reconciliation is the process of clarifying, verifying and reconciling medications on initiation or transfer of care, conducted in both ambulatory and inpatient settings2. The goal was to prepare medical students for their upcoming residency by teaching medication reconciliation during the hospital discharge process through a virtual session with didactic and interactive game material.

Methods: The 1.5-hour session occurred virtually as part of the 4th year medical student 2021 Transition to Residency course. It included didactics on medication reconciliation and cost of medications, with 5 game cases. The game was Electronic Health Record (EHR)-agnostic, web-based and interactive, with clinical cases for users to perform simulated medication reconciliation tasks with feedback on the patients’ outcome based on participant actions. Participants completed a total of 5 game cases over two 15-minute blocks, followed by two 5-minute debriefs. Pre- and post-session surveys were analyzed using chi-square in Python to evaluate change in student knowledge and confidence. Game usability was assessed using a modified version of the System Usability Scale (SUS). Trainee performance was evaluated using the game score for 2 cases pre- and post-didactic.

Results: Survey responses rates of 72/78 (92%) pre-didactic and 50/78 (64%) post-didactic. There was improvement in student quiz answers and self-reported knowledge and confidence. The average SUS score was 71 (good) and 94% of participants agreed that they would recommend the game in the future. Student performance on game cases did not improve pre- and post-didactic.

Discussion: During the pandemic, the need for social distance made it infeasible to have in-person teaching sessions for medication reconciliation. To maintain the hands-on simulation education, the mixed game and didactic session was feasible and scalable. Knowledge and confidence in medication reconciliation increased, and students responded favorably to the use of games in learning the material. Student game scores did not improve in the cases completed before and after the didactic. Each case tested a different clinical pitfall, and the game was not designed to measure improvement for each task. Incorporation of games to perform hands-on tasks has the potential to be practical and effective in delivering virtual clinical education.

References
Adoption of Telemedicine During the COVID-19 Pandemic: Perspectives of Primary Healthcare Providers

Joe Lintz, Doctor of Health Administration (DHA), Parker University, Dallas, Texas

Introduction
As the pandemic has unfolded across an array of communities worldwide, telemedicine has been promoted and scaled up to attend to and reduce person-to-person transmission of COVID-19 in the U.S. and internationally. This study examined the major barriers to adoption of telemedicine among primary healthcare providers at the primary care clinic of North Texas during the pandemic. In addition, the research gleaned information about respondent characteristics vis-a-vis telemedicine use.

Methods
A self-administered questionnaire was mailed to 67 primary healthcare providers at the primary care clinic of North Texas, with a 70% percent response rate (n=47). The survey collected information on perceived barriers to telemedicine adoption and experience with telemedicine in the last 12 months. In additional to descriptive statistics, multiple logistic regression was carried out to determine characteristics related with telemedicine used in the last 12 months and chi-square test to examine the relationship between number of perceived barriers and reporting any telemedicine use.

Results
Table 1 reveals that the most frequently mentioned barriers to telehealth adoption were “cost of equipment” and “lack of reimbursement” (58%), followed by “inadequate technical infrastructure” and “confidentiality and security concerns” (19%). In addition, 14% of respondents cited “resistance from patients, administrator, and users” as obstacles. Last, only 3% of respondents identified “lack of usefulness of telemedicine” as an impediment. Findings from the multiple logistic regression revealed that respondents employed fewer than 5 years (p=0.003) and younger than 40 years of age (p=0.002) and working as subspecialists (p = .011) were more likely than their counterparts to report telemedicine use. Plus, the results of the chi-square test indicated that the number of perceived barriers to telemedicine usage was negatively associated with using telemedicine. Provided in Figure 1 is a visualization that manifesting that there was a strong negative relationship between the number of perceived obstructions and telemedicine utilization.

Table 1. Percent of Respondents Mentioning Barriers to Telemedicine Adoption

<table>
<thead>
<tr>
<th>Barrier</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cost of equipment</td>
<td>30</td>
</tr>
<tr>
<td>Lack of reimbursement</td>
<td>28</td>
</tr>
<tr>
<td>Inadequate technical infrastructure</td>
<td>10</td>
</tr>
<tr>
<td>Confidentiality and security concerns</td>
<td>9</td>
</tr>
<tr>
<td>Lack of technical expertise</td>
<td>6</td>
</tr>
<tr>
<td>Resistance from patients</td>
<td>7</td>
</tr>
<tr>
<td>Resistance from administrators</td>
<td>4</td>
</tr>
<tr>
<td>Resistance from users</td>
<td>3</td>
</tr>
<tr>
<td>Lack of usefulness of telemedicine</td>
<td>3</td>
</tr>
<tr>
<td>Total</td>
<td>100</td>
</tr>
</tbody>
</table>

![Figure 1. Relationship Between the Number of Perceived Barriers and Telemedicine Use (n =47).](image)

Conclusion
The findings suggested that the barriers to using telemedicine existed across an array of situations and that decreasing these obstacles would be critical in encouraging future telemedicine adoption among providers during and after the pandemic. This will especially be the case for primary care practices where scarce financial resources have been a traditional problem for such providers. This complicacy is likely to be amplified owing to the inimical effects of the pandemic, during and after its potential containment or successful management of it.

Reference

Acknowledgements: This research was supported by Primary Care Clinic of North Texas Dallas, TX
Binary and Categorical Models for Automatic Acute Respiratory Distress Syndrome Detection

Kirill Lipatov, MD.1; Quan T. Do, PhD.2; Bradley J. Erickson, MD.3; Brian W. Pickering, MD.1; Vitaly Herasevich, MD., PhD.1

1Department of Medicine, 2Department of Anesthesiology and Perioperative Medicine, 3Department of Radiology, Mayo Clinic, Rochester, MN

Abstract
Early detection of Acute Respiratory Distress Syndrome (ARDS) is crucial to timely institute appropriate care and prevent adverse outcomes. The recognition relies in part on chest imaging interpretation and is hindered by delays and lack of inter-operator reliability. We derive and validate a Convolutional Neural Network (CNN) model for rapid automatic ARDS detection resulted in combined sensitivity of “probable ARDS” of 97% and specificity of 80% for “unlikely ARDS” categories.

Introduction
Acute respiratory distress syndrome (ARDS) is one of causes of respiratory failure associated with high mortality[1]. Timely recognition and institution of safe mechanical ventilation are crucial to avoid mortality. We sought to create a Convolutional Neural Networks (CNN) model that would automatically and rapidly detect ARDS with high specificity. We further combine the model with previously derived and validated ALI sniffer to improve the performance.

Methods
A cohort of 34826 chest x-ray images (15,899 unique patient admissions) were obtained, of which 3011 were used for training and testing (1002 “ARDS”, 1002 “Pneumonia”, and 1007 “Normal”). The images were used in training and testing the model with an 80% and 20% split respectively. The input size of images was set as [229x229x3]. This is the optimum input size for X-ray classification task [3,4]. Data augmentation methods were employed to avoid overfitting and therefore increases the model’s generalization ability. Convolutional Neural Network (CNN) was used to build models for “ARDS” and “Pneumonia” detection, which included 5 binary (“ARDS-Normal”, “Pneumonia-Normal”, “ARDS-Pneumonia”, “ARDS-NonARDS”) and one categorical classifications (“ARDS-Pneumonia-Normal”). NonARDS included combination of “Pneumonia” and “Normal” images.

Results
The categorical model (ARDS-Pneumonia-Normal) achieved the accuracy rate of 96%, with a testing accuracy rate of 86 %. The AUC (Area Under ROC Curve) of ARDS was 0.88, Normal 0.89, and Pneumonia 0.83. The True Positive (TP) rate of ARDS was 85%, while the TP rate of Normal and the TP rate of Pneumonia were 84% and 79% respectively. Model had Positive Predictive Value (PPV) of 0.84 with model recall of 0.83 and Negative Predictive Value (NPV) of 0.93. The results of binary classification are presented in Table II.

The validation sample included a separate cohort of ARDS patients used in derivation of the clinical sniffer and included 7,164 xray. The CNN model correctly identified 722 of 1046 images as “ARDS” and 3320 out of 3972 samples as “Non-ARDS”, resulting in overall sensitivity of 69% and specificity of 84% (PPV and NPV of 47% and 99% respectively). We further combined the more sensitive clinical parameters of the sniffer with more specific CNN-based image detection into the three-category classification: “probable ARDS” (both clinical parameters and CNN-model positive), “possible ARDS” (one of the two is positive), or “unlikely ARDS” (both negative). This resulted in combined sensitivity of “probable ARDS” of 97% and specificity of 80% for “unlikely ARDS” categories (Table III).

Conclusion
The automatic chest x-ray recognition models proved to have clinically useful performance that could accelerate detection of life-threatening respiratory conditions.
Survival Analysis of Breast Cancer from DNA Methylation Data

Guanghui Liu¹,²
¹School of Information Engineering, Nanjing University of Finance & Economics, Nanjing, China; ²Biomedical and Health Informatics, State University of New York at Oswego, Syracuse, New York, USA

Abstract

In this study, we propose a novel method for survival prediction using bidirectional LSTM network and ordinal Cox model. First, gene methylation expression data and clinical data are merged and filtered. Then, we build a deep LSTM model to predict patient survival risk. We use the leave-one-out method for cross validation and the concordance index to evaluate the prediction performance. Stringent cross-validation tests on the benchmark dataset demonstrates the efficacy of the proposed method.

Introduction

The Breast cancer is one of the most common form of dis-eases worldwide. It is reported that more than forty thousand women and four hundred men in the United States died from breast cancer annually before 2016. These data emphasize the importance of most profound understanding of factors that trigger breast cancer and contribute to its development. Gene methylation influence on cancer has been introduced with great success. Methylation of CpG sites is an epigenetic regulator of gene expression that usually results in gene silencing. Consequently, to explore the utility of methylation analysis for cancer diagnosis, we analyzed gene methylation expression of tumors from patients with breast cancer to identify potential cancer-specific methylation genes.

One long-term goal of cancer research is to identify prognostic factors that affect patients’ survival time, which in turn allows clinicians to make early decision on treatment. patients classified into a high-risk group may benefit from closer follow-up, more aggressive therapies, and advanced care planning. Cox proportional hazard¹ is among the most popular survival prediction models. Recently, based on the Cox model, several regularization methods have been proposed in the literature, e.g, LASSO-COX, RSF-COX, MLP, and DeepSurv et al. The above methods assumed that the survival information of one patient is independent from another, and thus miss the strong ordinal relationship between the survival times of different patients. Motivated by all these considerations, we thus propose a method with LSTM networks and ordinal Cox model to predict breast cancer patient's survival risk from gene methylation data.

Materials and Method

The methylation data was downloaded from TCGA² database. The benchmark dataset includes 779 patients with 20106 genes. Our proposed method has three stages, including the gene co-expression cluster stage, the bidirectional LSTM network stage, and the COX model stage. In the gene co-expression clustering stage, gene methylation expression data could be reduced in terms of feature dimension. WGCNA algorithm is used to cluster genes. So, twelve eigengenes are obtained and will serve as input features for the machine learning network. Secondly, the bidirectional Long-Short-Term-Memory (biLSTM) method is proposed to predict patient survival risk. Finally, a novel ranking loss function for the deep cox proportional hazard model is built for survival analysis to ensure that the ordinal relationship among the survival time of different patients can be preserved.

Conclusion

We assess the performance of the proposed method and carry out experiments on the training set through leave-one-out cross validation. The median risk score predicted by the cox proportional hazards model is used as a threshold to split patients into low-risk and high-risk groups. We test if these two groups have distinct survival outcomes using Kaplan-Meier estimator and rank test. The survival curves are drawn by applying different methods. Experimental results demonstrate the superiority of the proposed method over the existing four predictors. The good performances of the proposed method come from the use of the combined bidirectional LSTM predictor and ordinal information.

References

YouTube Video Analytics: Understandability Assessment of Diabetes Videos for Patient Education

Xiao Liu, PhD1, Anjana Susarla, PhD2, Rema Padman, PhD3
1Arizona State University, Tempe, AZ; 2Michigan State University, Lansing, MI; 3Carnegie Mellon University, Pittsburgh, PA

Introduction
Digital therapeutics that leverage mobile applications, software platforms and multi-media content have been developed for the prevention and/or management of a variety of chronic and acute health conditions, such as diabetes1. However, enhancing the effectiveness of patient education and health communication is essential to achieve improved uptake of such therapies for better health outcomes. In this study, we draw on the validated Patient Education Materials Assessment Tool (PEMAT)2 to assess understandability of publicly available user generated social media content such as diabetes-related YouTube videos.

Methods
Building on prior research1 and guided by PEMAT2, we develop a scalable, generalizable, augmented-intelligence and co-training-based machine learning approach to retrieve and assess the understandability of encoded medical content in YouTube videos on diabetes-related topics. Over 200 search keywords related to diabetes patient education are identified using literature review and clinician inputs. We collected the top 50 videos for each search term using YouTube Data API resulting in 9,873 unique videos for the study, and extract their content and metadata. Two clinicians evaluated 700 randomly sampled videos from the corpus for video understandability according to the PEMAT guideline, with inter-rater reliability assessment providing a consensus score, and used for training the classification model. We define the classification of video understandability for patient education as a multi-view learning and binary classification problem using a co-training machine learning model3. Medical terms are extracted and analyzed using a deep learning model1. We utilize the Google Cloud Video Intelligence platform to perform video data processing and extract features that are relevant for understandability. We perform object detection to obtain labels of objects in the videos at frame level and detect scene changes within the videos to examine whether videos are chunked into distinct segments. We use speech transcription to understand the narratives and extract features from them to evaluate speech clarity, and optical character recognition to detect text, tables and illustrations in the video. The confidence levels associated with these methods provide guidance on the quality of the data processing methods. All these features are subsequently used by the co-training model to assess understandability.

Results
The co-training model with logistic regression classifier initially starts with 600 labeled videos for supervision. After 12 iterations and adding 305 video annotations, the model converged, and is evaluated on 100 held-out labeled videos. Our approach achieved a precision of 0.84, a recall of 0.79, and an F1 score of 0.81 in classifying videos with high understandability. A comparison with three benchmark models, logistic regression, Support Vector Machines, and Random Forest, showed significantly improved video understandability classification by the co-training model. Furthermore, measuring average precision at K, K=1...10, for 20 random queries, a medical expert ranked 72% of top 10 videos recommended by the co-training approach compared to 40% of those based on YouTube’s default ranking.

Conclusions
In this study, we develop an automated, scalable and multi-modal algorithmic solution to evaluate whether medical information encoded in diabetes-related YouTube videos adheres to validated understandability guidelines on patient education and health promotion, and retrieve videos that meet these criteria. Ongoing research will extend the current methodologies to include critical criteria such as accuracy, actionability and quality of the video content and eventually create a repository of curated health videos. We anticipate that this approach can facilitate recommendations of highly ranked and clinically validated YouTube videos to information seeking consumers and point-of-care instructions by clinicians for improved patient education, self-care management and societal health literacy.

References

1730
Engaging the disengaged: recommendations for partnering with transitional aged youth to improve mental health help-seeking in the digital age

Brian Lo, MHI\(^1\), Jenny Shi, MPH\(^1\), Howard W. Wong, MHI\(^1\), Alexxa Abi-Jaoudé, MPH\(^1\), Elisa Hollenberg, MSW RSW\(^1\), Andrew Johnson, BA\(^1\), David Wiljer, PhD\(^{1,2}\)

\(^1\)Centre for Addiction and Mental Health, Toronto, ON, Canada; \(^2\)University of Toronto, Toronto, ON, Canada

Introduction

Digital health tools, such as mobile apps, can help support transition-aged youth (TAY) between 16-29 years old navigate challenges related to finding mental health and wellness resources in Canada. Co-design approaches have been suggested to help support the development of digital health tools that meet the needs of end-users \(^1\). However, there is a lack of guidance on how to effectively engage with TAY throughout the development and optimization of these platforms. This presentation aims to share the lessons learned from engaging TAY in the development of Thought Spot, a mobile health app designed to support TAY in seeking mental health and wellness resources \(^2\).

Methods

In 2016-2019, we conducted 5 co-design workshops with TAY to develop the Thought Spot platform \(^2\). As a part of the solution development, we also explored recruitment and engagement approaches for TAY through 12 focus groups\(^3\). Based on a review of the academic literature and a reflection of our challenges and lessons learned from the workshops, we have developed seven recommendations for engaging TAY in digital health initiatives.

Results

From the activities that were conducted for the development and optimization of Thought Spot, several recommendations were identified and are described in Table 1 below.

<table>
<thead>
<tr>
<th>Recommendations</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Address power dynamics with transparency and clear boundaries</td>
<td>Clearly establishing the scope and decision-making capacity of TAY can ensure that their perspectives are valued and issues related to power dynamics are mitigated</td>
</tr>
<tr>
<td>Language plays a key role</td>
<td>Use of non-stigmatizing language and avoiding colloquial language (e.g., the term ‘research’) can reduce stigma and language barriers</td>
</tr>
<tr>
<td>Account for diversity and foster a culture of respect</td>
<td>Project activities that are considerate of sociocultural influences can help foster inclusivity among a diverse group of participants</td>
</tr>
<tr>
<td>Involve TAY in all aspects of the project</td>
<td>TAY should be involved from planning to evaluation of all project activities to ensure appropriateness and suitability</td>
</tr>
<tr>
<td>Build upon relationships with key members and leaders of the TAY communities</td>
<td>Key community members can help researchers develop the trust and relationships needed for recruitment and engagement activities</td>
</tr>
<tr>
<td>Consider and address barriers to engagement (e.g., resources)</td>
<td>Identifying and exploring challenges that hinder engagement (e.g., time and resources) can reduce barriers for TAY to participate</td>
</tr>
<tr>
<td>Be flexible and innovative to emerging approaches favored by TAY</td>
<td>Use of emerging approaches (e.g., social media platforms) in creative ways is critical for engagement and recruitment of TAY</td>
</tr>
</tbody>
</table>

Conclusion

A set of recommendations were identified for engaging TAY in the development of a digital mental health tool. Future studies should verify these recommendations through other digital health projects for TAY.

References

Monitoring the Effects of the Pandemic on Cancer-Related Patient Encounters

Jack W. London, PhD¹; Elnara Fazio-Eynullayeva, MA²; Matvey B. Palchuk, MD, MS²; Peter Sankey, MBChB³; and Christopher McNair, PhD¹
¹Thomas Jefferson University, Philadelphia, PA, ²TriNetX LLC, Cambridge, MA, ³University Hospitals Plymouth NHS Trust, Plymouth, United Kingdom

Introduction

We developed an analytical infrastructure to quantify the immediate impact COVID-19 has had on deviation from normal cancer care activities, including cancer screening efforts. We leveraged an existing health research network platform (TriNetX; Cambridge, MA)¹ to analyze data from over 20 different provider institutions across the United States that represent 28 million patients and that have immediate data on relevant patient encounters. We focused on patients seen at these healthcare institutions with diagnostic codes for malignant, benign, in situ, and unspecified neoplasms. We initially looked at the number of these patients having encounters for March and April 2020 compared with the same months in 2019. Subsequently, we repeated similar analyses for the rest of 2020. These analyses include patients being seen for the first time at these institutions, possibly for screening, initial diagnosis, second opinion, or treatment initiation. Finally, patient metrics for site-specific malignancies were analyzed (lung, breast, prostate, colorectal, hematologic cancers and melanoma). Furthermore, we considered the data from a healthcare institution in the United Kingdom to initiate exploration of the global applicability of the trends we observed for our US network.

Results and Discussion

Development of this analytical infrastructure began in April 2020. The first published results² for this early period of the pandemic identified significant declines in encounters for both existing and new incidence neoplasms as well as substantial decreases in breast cancer and colorectal cancer screenings. These analyses were repeated over subsequent periods of time as the pandemic progressed through 2020, and a recovery in the numbers of cancer screenings and newly diagnosed cancers were observed.

Conclusion

Our findings indicate an initial significant decline in new incidence neoplasm-related encounters and cancer screenings as a result of the COVID-19 pandemic into April 2020, which suggests the potential for increases in later-stage disease at initial presentation and increased screening demand in the future. This decline was followed by an increase in new incidence encounters during the subsequent weeks, although 2019 levels were not quite reached. As expected, mammography and colorectal screenings did surpass prior year numbers later in 2020. Together, this confirms the need to monitor both the short-term and the long-term effects the pandemic will have on this population. The analytical infrastructure described here provides the means for this surveillance.

References

Oral-HPV Screening Tool for Assessing Individuals at Risk for HPV-associated Oropharyngeal Cancer

Steven D. London, DDS, PhD, MPH, MBA 1,2,3, Paul Fontelo, MD, MPH 1, Shadi Boroumand, DMD, MPH 2, Bruce A. Dye, DDS, MPH 2.

1National Library of Medicine, Bethesda, Maryland
2National Institute of Dental and Craniofacial Research, Bethesda, Maryland
3Stony Brook University School of Dental Medicine, Stony Brook, New York

Introduction: Human papillomavirus (HPV) is the most commonly transmitted sexual infection in the US and is associated with 70% of all oropharyngeal squamous cell cancer (OPC) in the US leading to significant morbidity, mortality, and cost. At the present time preventive interventions are focused on vaccinating teenagers and those up to age 26, but interventions for adults still rely on screening through clinical oral cancer examination and counseling. However, it is important for oral health professionals to inform patients of risk and screen their patients for early OPC detection. D’Souza, McNeel and Fakhry developed an algorithm for determining risk for oral HPV infections. The algorithm assigns risk based on sex, number of lifetime oral sex partners, and current smoking status. We further adapted the algorithm to include an indicator for the number of remaining natural teeth. Previous research has indicated that significant tooth loss is associated with increased likelihood of having oral HPV even after controlling for the influence of smoking. Risk for oral HPV varies for men and women, dependent upon the number of oral sex partners. For men who currently smoke, risk is even higher. Risk is also elevated for men and women with 12 or fewer teeth who do not smoke. The goal of this project is to ascertain the potential impact of oral HPV screening performed by oral health professionals using this brief screening tool in the US adult population if practice guidelines were implemented that included the use of this screening tool.

Methodology: The aim of this study is to quantify the impact of oral health professionals who provide screening for at-risk individuals using the screening tool. Data from 9183 individuals (20-59 years old) participating in NHANES 2011-2016 with valid information on age, sex, smoking status, self-reported sexual behavior information, number of remaining teeth and laboratory-assessed oral HPV status, were used. Participants were categorized into one of the six risk categories based on sex, smoking status, and having 12 or fewer teeth. Regression analysis determined that men are not at very low risk and women are not at elevated risk. The six groups are: low risk men, medium risk men, and elevated risk men; very low risk women, low risk women, and medium risk women. These six risk groups were further subdivided into three groups to reflect risk status as low, medium, and elevated for oral HPV infection. Analysis was performed using SAS statistical software 9.4 analytic routines for frequency calculations and test performance characteristics.

Results: Approximately 7% of the US population were HPV+ (10.2M) in 2011-2016. After sorting approximately 144M Americans into the 3 main risk categories using the modified screening guide, we estimate that 109M Americans would be identified as low risk (75.6%), 25.4M as medium risk (17.6%), and 9.8M (6.8%) as having elevated risk for HPV-associated oropharyngeal cancer. Among Americans, 83.5M who had dental visits in the past 12 months, 65M would have screened as low risk, 14.2M as medium risk, and 4.3M as elevated risk. Among these dental care users, 5.5M were oral HPV+, whereas 4.7M non-users were HPV+. Additionally, among those with a dental visit in the past 12 months, 23.4% received oral cancer screening advice. Among those not receiving advice, 16.4% would have screened as medium or elevated risk. If all individuals who visited the dentist had received advice and the oral HPV screening tool was used, an additional 63.9M Americans would have been screened identifying 13.4M as either medium or elevated risk for oral HPV+. Among these individuals, 3M of the 5.4M HPV+ dental users could have been referred for oral HPV testing. The performance of this screening tool yielded a sensitivity of 0.56 and a specificity of 0.78, with a 76% accuracy.

Conclusion: Findings from these analyses indicate that clinical risk screening for oral HPV could quickly identify a substantial number of dental care users who are medium to elevated risk for oral HPV (13.4M American adults) as well as more than half of all HPV+ dental care users (3M adults). Because 70% of all Oropharyngeal cancers are associated with HPV, better identification of dental patients who are medium-elevated risk for oral HPV should result in improved patient advising, referring, and decision-making with the goal of reducing OPC incidence and mortality in the US. We suggest that the screening tool could be integrated into an ICT tool for clinical use at some time in the future.

References


Identifying Criteria for Antonym Generation from Corpora

Chris J. Lu, PhD1,2, Amanda Payne, PhD1,2 and James G. Mork, MSc1
1National Library of Medicine, Bethesda, MD 2Medical Science & Computing, LLC, Rockville, MD

Introduction
The SPECIALIST Lexicon (the Lexicon) and Lexical Tools are enhanced with new antonym features for the 2022 release [1]. Antonym pairs (aPairs) are generated from the Lexicon, suffix and prefix derivations, collocates (CC) and semantic relations (SC) in corpora [2]. Criteria are identified from a set of common aPairs to effectively generate antonyms from CC and SC. This paper describes our systematic approach on identifying generic criteria for antonym generation from corpora.

Methods
We collected commonly used antonyms from 14 sources on the internet [3]. This collection includes 1000 unique, lowercased, single word aPairs, which are assumed to have representative characteristics of overall antonyms and is used as a training and test set (TtSet) to identify generic criteria of antonyms. APairs in the TtSet are manually tagged for canonicity, domain, type, and negation. Canonical aPairs have a generic domain, that is central to human life and ways of living across time and cultures. Computer programs are developed to 1) retrieve properties of aPairs, such as EUIs, POSs, CUIs, STIs, sources, etc. 2) compute stats among properties to identify generic criteria of antonyms. TtSet and 2021 antonym production data are used for this study.

Results – Identified Criteria
First, aPairs from TtSet and 2021’s data set have the same (10) domains for canonical antonyms and similar negation rates (9.51% and 7.14%, respectively). This conforms with our hypothesis that TtSet is representative of overall antonyms to retrieve antonym characteristics. Second, Table 1 shows the source distribution of TtSet. We observed antonyms from corpora models (CC and SC) to be worthy of further development because they contain the most aPairs (83.65%). Third, canonical aPairs must be in the Lexicon with valid EUIs and have the same POS. Fourth, canonical aPairs cannot be synonyms. This confirms the theory that antonyms and synonyms are similar in domain and different in polarity. Fifth, antonyms should have CUIs (our scope is using concepts in the UMLS-Metathesaurus) and share STIs because 67.79% of aPairs share STIs when they have CUIs. The analysis’ results are shown in Table 2 and used to retrieve antonym candidates from corpora. Our ultimate objective is to provide generic and comprehensive antonym features with completion of antonym generation from corpora models. The Lexicon is distributed with UMLS by NLM via an Open Source License agreement and is available at: https://umlslex.nlm.nih.gov/lexicon.

Table 1. Source distribution of canonical aPairs in the TtSet.

<table>
<thead>
<tr>
<th>LEXICON</th>
<th>SuffixD</th>
<th>PrefixD</th>
<th>CC</th>
<th>SC</th>
</tr>
</thead>
</table>
| TtSet canonical aPairs | 1.95% | 0.58% | 13.81% | 33.07% | 50.58%

Table 2. Stats of analyzed properties for canonical aPairs in the TtSet.

<table>
<thead>
<tr>
<th>EUIs</th>
<th>Same POS</th>
<th>Synonyms</th>
<th>Share STIs (if have CUIs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TtSet canonical aPairs</td>
<td>100%</td>
<td>100%</td>
<td>0%</td>
</tr>
</tbody>
</table>

Acknowledgements
This research was carried out by staff of the National Library of Medicine (NLM), National Institutes of Health, with support from NLM.

References
An exploration of information extraction models on transcribed patient visits

Kevin Lybarger, PhD, Erica Qiao, BS, Meliha Yetisgen, PhD
University of Washington, Seattle, WA, USA

Introduction

Documentation burnout contributes to clinician job dissatisfaction, and clinical notes often omit salient information. The automatic generation of notes from clinician-patient conversations, referred to here as ambient note creation (ANC), could provide an alternative, time-saving documentation process. ANC involves reorganizing dialogue content by note sections and translating the conversation to medical language. Although ANC has been pursued as a sequence-to-sequence machine translation problem, ANC may benefit from incorporating information extraction (IE) models that identify salient clinical information (e.g. signs, symptoms, treatments) and create abstractive summaries. In this work, we analyze transcribed conversations and evaluate the performance of IE models designed for clinical notes.

Methods

In a study of physician communication skills and training interventions, 1,282 primary care visits were audio recorded at clinics around Seattle, WA. The dialogue from a subset of these recordings was manually transcribed, including 144 visits at the University of Washington. We use these transcripts to explore the dialogue contents and the limitations of existing IE systems. In our previous work, we developed high performing IE models trained on clinical notes for characterizing social determinants of health (SDOH) and symptoms. The SDOH include alcohol, drug, and tobacco use, employment status, and living situation. Here, we annotate 10 randomly selected transcripts using the same SDOH and symptom annotation schemes and evaluate the performance of these extractors in this conversational domain.

Results and Discussion

Table 1 summarizes the turns, words, and questions per visit. The length of the patient and doctor contributions (words per visit) is similar, with patients making more frequent, shorter contributions. Medical assistant (MA) contributions are relatively small. In the annotated transcripts, there are 12.2 symptoms per transcript. On clinical notes, the symptom extractor identified symptoms at 0.83 F1 and predicted assertion values (present vs. absent) at 0.79 F1; however, the performance drops markedly on the dialogues to 0.64 F1 for symptom identification and 0.45 F1 for assertion. The more colloquial language of the dialogues, relative to clinical notes, is a contributing factor to the decreased symptom identification performance (e.g. “I feel crappy”). Symptom identification errors result in cascading assertion errors; however, the decreased assertion performance is also associated with the frequent coreferences to symptoms across multiple sentences and turns. The annotated transcripts only include 5 substance use and 4 employment events; however the SDOH extractor predicted 58 substance use and 18 employment events (performance < 0.04 F1). False positives related to substance use include the mislabeling of prescription drugs as illicit drugs, as well as difficult to explain errors. Many of the employment-related false positives are associated with phrases, like “works for me,” where “work” is used outside the context of employment.

ANC has the potential to reduce documentation burden and improve documentation quality. It may benefit from IE models that identify critical information. Our findings indicate existing IE models do not generalize well to conversational data. The degraded performance would be exacerbated by speech recognition errors associated with the automatic transcription of the dialogues. In future work, we plan to explore IE in this clinical conversational domain, with the goals of improving extraction performance and integrating the extracted information into an ANC pipeline.

References

The Real-World Trend Analysis of Cancer Drugs Empowered by Natural Language Processing

Meng Ma¹, Kyeryoung Lee¹, Yun Mai¹, Christopher Gilman¹, Zongzhi Liu¹, Mingwei Zhang¹, Minghao Li¹, Arielle Redfern¹, Tommy Mullaney¹, Tony Prentice¹, Paul McDonagh¹, Qi Pan¹, William Oh¹,², Rong Chen¹,², Eric Schadt¹,², Xiaoyan Wang¹
¹Sema4, Stamford, CT; ²Icahn School of Medicine at Mount Sinai, New York, NY

Introduction
Electronic health records (EHR) have been widely used in clinical practice and have generated a vast amount of clinical data, which has become a rich resource for quality of care and clinical research. Investigation of the trend of cancer drugs in clinical practice is an important step for the evaluation of the real-world performance of cancer drugs. This study applied natural language processing (NLP) to recognize cancer treatments from clinical notes of lung cancer patients and investigated the trend of cancer drug usage for lung cancer treatment.

Methods
The lung cancer cohort of patients was curated from the Mount Sinai Data Warehouse, and 461,098 clinical notes were extracted for this cohort (2010.01-2019.12). We built a Bidirectional Long Short Term and Conditional Random Fields (BiLSTM-CRF) based NLP model on recognizing cancer treatments. The resulting model was validated independently with an F1 score of 95%. All distinct cancer treatments recognized from the 461,098 notes by the pipeline were normalized and post-processed with a manual review (Figure 1). Time-trend analysis of the drug usage in lung cancer patients was investigated in this study.

Results
The NLP pipeline identified 2,891 more patients with detailed cancer treatments compared with the total of 1,395 patients from structured data. The approval history and longitudinal use of the most frequently appearing cancer drugs (chemotherapy, targeted therapy, and immunotherapy) are shown in Figure 2. The continuously increasing amplitude of chemotherapy use is far beyond targeted therapy and immunotherapy drugs, implying that although targeted therapy and immunotherapy have revolutionized the landscape of lung cancer therapeutics, chemotherapy remains an essential component. Compared to the first and second-generation EGFR TKI Erlotinib and Afatinib, the third-generation EGFR TKI Osimertinib shows a dramatic increase in clinical use since FDA approval in 2015. The anti-VEGF monoclonal antibody Bevacizumab is the second most frequent targeted therapy drug, while a slight decrease in use since 2017. Pembrolizumab use has rapidly grown since clinical trial KEYNOTE-024 succeeded in phase III in June 2016, while Nivolumab use has gradually decreased ever since clinical trial CheckMate-026 failed in phase III in August 2016. Atezolizumab and Durvalumab both showed an increasing trend of use in clinical practice. The data showed a consistent trend with the published revenue of the cancer drugs from structured and manually reviewed data.

Conclusion
This study demonstrates NLP is a powerful tool to enable real-world free-text clinical notes as an analyzable resource to facilitate the investigation of the use trends of cancer drugs in clinical practice.
Evaluating Social Determinants of Health in Clinical Communications Data

Diwakar Mahajan, MS¹, Jennifer J. Liang, MD¹, Ananya Poddar, MS¹, Sasha Ballen, MS², Ching-Huei Tsou, PhD¹
¹IBM TJ Watson Research Center, Yorktown Heights, NY; ²R-Health Inc, Elkins Park, PA

Introduction

Social determinants of health (SDOH) contribute to a patient’s functional status and disease onset and progression. With the shift towards telemedicine during the COVID-19 pandemic, use of secure messaging between patient and provider is becoming more prevalent and may prove to be a valuable source of SDOH information. Unlike other patient record data, communications data have the benefit of capturing the patient’s concerns in real-time in their own words. This work presents a preliminary analysis of SDOH in clinical communications data.

Methods

Our corpus consists of R-Health communications data from January 2019 to September 2020, containing a total of 672,725 messages over 11,186 patients. Starting from the core domains identified by the Institute of Medicine’s report on capturing social and behavioral determinants of health in EHRs¹, we (1) identified a subset of SDOH topics of interest: financial resource strain, stress, depression, physical activity, tobacco use and exposure, alcohol use, social connection/isolation, exposure to violence, (2) created a set of keywords for each SDOH topic, and (3) generated a keyword-based script for each SDOH topic to automatically detect topic-specific “chatter” or discussion in communications data. To evaluate this script, we annotated a total of 300 messages, half of which were randomly sampled, and the remaining selected using our keyword-based script to ensure sufficient positive instances. As the resulting dataset still contained <10 instances for financial resource strain, physical activity, alcohol use, social connection/isolation, and exposure to violence, those topics were excluded from analysis. Table 1 presents the ground truth counts and evaluation of the keyword-based script for the remaining SDOH topics. To uncover trends for different SDOH topics in communications data, we ran this script over our entire corpus of 672,725 messages.

Results

Figure 1 presents the percentage of patients with SDOH “chatter” broken down by topic and over time. Here we restrict our analysis to only those messages sent by the patient. As observed, there is a significant increase in patient-initiated messages about stress starting around March 2020 when COVID-19 was officially declared a global pandemic. The “chatter” peaked in April 2020 and has come down slightly but remained higher than the previous baseline. A similar but less dramatic trend can be observed for depression.

Conclusion

We studied presence of SDOH “chatter” in communications data for a direct primary care practice and observed interesting trends for discussions about stress in light of the current COVID-19 pandemic. In future, we plan to generate more ground truth, expand our analysis to other SDOH topics, and do a deeper analysis to better characterize the SDOH discussions in communications data.

References


Table 1. Evaluation of SDOH keyword-based script

<table>
<thead>
<tr>
<th>SDOH Topic</th>
<th>#Message</th>
<th>Micro F₁</th>
<th>Macro F₁</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stress</td>
<td>50</td>
<td>0.93</td>
<td>0.88</td>
</tr>
<tr>
<td>Depression</td>
<td>14</td>
<td>0.97</td>
<td>0.86</td>
</tr>
<tr>
<td>Tobacco use and exposure</td>
<td>11</td>
<td>0.99</td>
<td>0.93</td>
</tr>
</tbody>
</table>

Figure 1. Patient-initiated SDOH "chatter" over time
COVID-19 as a Catalyst: Rapid Development of Usable Digital Patient Services

Andrea N. Mahnke, MS1, Jenna R. Kautza1, Emily A. Wiedoff1, Daniel P. Burish1, John B. Welch, MA1
1Marshfield Clinic Health System, Marshfield WI

Objective

Overall, the objective of this initiative was to rapidly develop digital services in order to help keep patients safe and handle increased volumes during a global pandemic. These services included a mobile app, online patient appointing and aspects of telehealth appointments. An objective of iterative user testing in parallel with development was to determine where new digital services performed well and where they failed to meet the needs of patients. A second objective of this work is to inform further design enhancements and developments of above digital services.

Background

Planning for digital patient services was underway pre-COVID-19, but the onset of the pandemic rapidly accelerated development and implementation efforts. While the imperative was to get the digital services in the hands of patients quickly, the team wanted to ensure that the services would be usable for patients. Usability is a key factor in adoption of digital services, especially for patients with complex clinical landscapes and who may also represent underserved populations.1

Methods

Pre-COVID-19 two surveys were conducted with health system patients. One was to assess patient technology use in the health system’s rural setting. The second, used the Kano Model to assess patients’ desires for mobile app features in an effort to prioritize them for development. Starting in April of 2020, we conducted iterative testing on the online patient self-appointing feature. The usability testing of online patient self-appointing included a patient portal prototype and production platform. This was followed by testing of a mobile app prototype. We also conducted a satisfaction survey of online patient self-appointing (ongoing). A card sort was conducted to help understand how patients would organize different features of a mobile app. This was followed by usability testing of mobile app home/dashboard screens and mobile app global feature navigation with four prototypes. All testing iterations were conducted with health system patients and employed remote moderated task based usability evaluations, followed by semi structured interviews. When time permitted during the semi structured interviews, we asked questions regarding patients’ perspectives of telehealth appointments, whether they had participated in a telehealth appointment or not. An extensive telehealth survey was sent to all patients that had participated in a telehealth appointment and had an email on record.

Results

The technology use survey was completed by 316 patients. The survey provided details on smart phone use and broadband access across demographics. A total of 9,865 patients completed the Kano Model survey. This identified the app features that patients most expected to see. There were three iterations of online appointing evaluations with a total of 35 patients. Feedback from patients resulted in improvements of labeling and task flows. A total of 216 patients have completed the online appointing satisfaction survey. The survey is still open and analysis is in progress. The mobile navigation card sort was completed by nine patients and provided useful information, especially about how patients expect to view and group information. The mobile app home/dashboard screens and global app feature navigation was conducted with 24 patients. Valuable information was gained regarding placement of screen elements and navigation functionality, particularly regarding the menu. The telehealth survey was completed by 2,945 patients and analysis is in progress.

References

Quantifying Changes in Resident-Patient Interactions During the COVID-19 Pandemic Using EHR Audit Logs

Mark V. Mai, MD, MHS¹, Naveen Muthu, MD¹, Bryn Carroll, MD¹, Anna Costello, MD¹, Daniel C. West, MD¹, Adam C. Dziorny, MD, PhD²

¹The Children’s Hospital of Philadelphia, Philadelphia, PA; ²University of Rochester School of Medicine, Rochester, NY

Problem: Curricular adaptations to limit trainees’ exposure to COVID-19, in addition to changes in healthcare utilization due to the pandemic, likely affected trainees’ participation in direct patient care. Missed clinical experiences will potentially have longstanding effects on the quality of the medical workforce. Recent methods of patient attribution, using EHR audit logs, can inform educators about trainee-patient experiences.

Methods: We performed a retrospective cohort study with year-over-year analysis of pediatric resident provider-patient interactions (rPPI) at a large academic freestanding children’s hospital. We examined cohorts of pediatrics residents during March 15 – June 15 in 2020 (Exposure), 2018 (Control #1) and 2019 (Control #2). Using previously described methods of patient attribution from EHR data, we extracted and analyzed resident shifts and rPPIs.

Results: Residents in the Exposure period had fewer shift counts than those in both of the Control periods, across all PGYs (p < 0.05). PGY1 residents in the Exposure cohort experienced the largest absolute decrease in rPPIs across all care settings (Exposure: 111, Control #1: 222, Control #2: 206), particularly in the Continuity Clinic context (Exposure 8.2, Control #1: 36.7, Control #2: 36.5), where patient continuity also decreased from 23.4% to 11.1%. Common AHRQ Clinical Classifications Software Refined (CCSR) diagnosis categories previously encountered by pediatric residents in clinic decreased during the Exposure period (Figure 1) with similar findings in ED encounters.

Conclusion: While necessary to protect trainees, curricular changes coupled with changes in healthcare utilization drastically affected clinical learning opportunities through authentic patient interactions. Patient attribution algorithms can facilitate analysis of external disruptors to traditional models of medical education, allowing educators to better support learners’ needs and development, as well as plan for potential effects on competencies of groups of trainees.

References
Public Perspectives on the Ethical Collection and Sharing of Consumer-Generated Health Information

Sabrina Mangal, PhD, RN1, Leslie Park, BS1, Meghan Reading Turchioe, PhD, MPH, RN1, Lisa Grossman Liu, PhD2, Annie C. Myers, MA1, Brittany Taylor, BS, RN2, Parag Goyal, MD, MSc3, Lydia Dugdale, MD, MAR5, Ruth M. Masterson Creber, PhD, MSc, RN1

1Department of Population Health Sciences, Division of Health Informatics, Weill Cornell Medicine, New York, NY; 2Department of Biomedical Informatics, College of Physicians and Surgeons, Columbia University, New York, NY; 3Columbia University School of Nursing, New York, NY; 4Department of Medicine, Divisions of Cardiology and General Internal Medicine, Weill Cornell Medicine, New York, NY; 5Department of Internal Medicine, Columbia University, New York, NY

Introduction

As patient-centered care continues to grow as a cornerstone to clinical care delivery, the implementation and assessment of patient-reported outcomes (PROs) has become an area of interest for clinicians to holistically understand patients’ health statuses. Since PROs provide valuable insights into patients’ clinical pictures (e.g. symptom information, quality of life), analysis of PROs has become a common source of research studies and data analyses that aim to refine or further develop patient-centered interventions. However, recent legislation encourages increased transparency of individual health information to patients and more efforts to engage patients in their care. Part of this transparency involves the question of whether or not researchers are obligated to return this collected health information to participants to encourage transparency. Moreover, public perspectives on trust in scientific research are varied and lacking clinical context, which may influence preferences for data collection and sharing in research. Therefore, in this study, we aim to explore public perspectives on collecting, returning, and sharing their health information, and how different use cases might change their trust in research.

Methods

Guided by the Ethical Framework of Consumer-Generated Data and the Socio-Ecological Model, we aimed to conduct two surveys: one with a representative US sample, and one with a representative sample of New York State (Empire State Poll). In these surveys, we explore: 1) trust in research compared to trust in the general public, 2) preferences for researchers to return collected health information to participants, 3) desired format to receive health information, and 4) the effects on trust in research based on multiple data-use cases (based on use cases, participants rated their trust on a scale from 1 (strongly decrease trust) to 5 (strongly increase trust)).

Results

Overall, participants to date (n = 502, anticipated n = 1,300) had an average age of 46.6 years and were 51% female, 74% White, and 7% were of Hispanic or Latino origin. Most participants (66%) attended some college or had a bachelor’s degree, and 28% had limited/marginal health literacy. At baseline, most participants (60%) trust scientific researchers more than the general public. When given a scenario where health information was collected for research purposes, 84% of participants wanted access to their health information. On a scale of 1 (not important) to 10 (extremely important), most participants found it important to have access to the health information that researchers collect about them (avg: 8.0/10.0). Participants generally preferred to see their health information in E-mail (31%) or online website (17%) format. Many (85%) respondents indicated that having access to their collected health information would increase trust in research, and when knowing that the information would improve research about their own health condition, 70% of respondents indicated that their trust in research would increase.

Conclusions

It is important for participants to have access to their collected health information. Generally, participant trust in research is high when the purpose and implications of using and sharing their health information are disclosed. In future research studies, researchers should consider returning information back to participants and focus on transparency in disclosing the intended purpose and uses of their shared health information to foster trust.

References


This research is supported by the National Institute of Nursing Research. (R00NR016275; R00NR016275-05S1; PI: Masterson Creber)
PainRE-Life: A FHIR Based HUB for the management and support of patients with chronic pain

Sara Marceglia¹, Vania Manzelli², Pierluigi D’Antrassi³, Alfredo Cuzzocrea⁴, Annamaria Caruso¹, Marco Prenassi¹, Chiara Savino², Maria Teresa DePippo², Stefano Lenzì⁵, Roberta Ferrucci⁶, Costanza Conti³, Giulia Candiani⁷, Francesca Memini⁷, Elda Judica⁸, Massimo Corbo⁹, Marianna Masiero⁶, Gabriella Pravettoni⁶,⁹

¹Dip. Ingegneria e Architettura, Università degli Studi di Trieste, Italy; ²Nuvyta Srl, Italy; ³IMEA Lab, University of Calabria, Italy, Italy; ⁴Euleria Srl, Italy; ⁵Università degli Studi di Milano, Italy; ⁶Zadig Srl, Italy; ⁷Dept. of Neurorehabilitation, Casa di Cura del Policlinico, Italy; ⁸Applied Res. Division for Cognitive and Psychological Science, Istituto Europeo di Oncologia, Italy

Introduction

Chronic pain is a burdensome condition, with high social impact, with the patients experiencing a limited working capacity, psychiatric issues, and general disability[1]. Patient’s therapeutic pathways include multiple specialists, while continuity of care and care coordination are needed to guarantee pain management. Pain is therefore considered as one of the most promising conditions in which the use of digital health technologies, ecological momentary assessments, and digital communication tools may boost patient’s engagement and coping, as well as better monitoring of therapeutic outcomes[2]. PainRE-Life is a project, financed by the Lombardy Region (Italy), aimed to develop a dynamic and integrated technology ecosystem to allow care continuity in patients with pain.

System architecture

The project is based on three main technological assets (Figure 1): first, a FHIR based cloud platform (Nu Platform, developed by the partner Nuvyta) that enables care pathway definition and data collection; second, a big data infrastructure connected to the FHIR server that analyzes data and provides inputs to an ecosystem of personalized applications for the decision support of patients and caregivers (third asset, developed by the partner Zadig). Two clinical centers (Casa di Cura del Policlinico Rehabilitation Center, and IEO Oncology Hospital), and a home-based telerehabilitation service provider (Euleria Srl) will serve as case studies for the PainRE-Life platform. The project will also provide an e-learning platform for general practitioners (GPs). Finally, the system was also designed to be integrated, in the future, to routine healthcare, through electronic health records.

Results and Conclusions

To date, the system implements three digital care pathways for the management of pain in post-stroke, early breast cancer, and fibromyalgia (treated with transcranial direct current stimulation, tDCS) patients. Each care pathway involves the execution of a workflow in the Nu platform, that allows collecting all the data (pain severity scales, pain location, psychologic assessment) to be completed by the patient/caregiver at home or by the healthcare professional in the hospital setting and the integration with devices providing patient’s related information (tDCS devices and smartbands for activity monitoring). The validation will start in June 2021, and will first include usability testing and the verification of the care-pathway integration level, by monitoring the deviations from guidelines and recommendations. A second step will be the use of the platform for the evaluation of therapeutic outcomes.

In conclusion, the system will be developed to remain as a full service for the Lombardy Region, and as a reference for all the conditions that require continuity of care, multidisciplinary approaches, and heterogeneous data collection to effectively support patients and caregivers.

References

Quantifying Spatial Tumor Heterogeneity Using Single-Cell Proteomics

Adriano Martinelli, MSc¹,², Aditya Kashyap, PhD¹, Anna Fomitcheva Khartchenko, PhD¹, Govind Kaigala, PhD¹, Maria Anna Rapsomaniki, PhD¹
¹IBM Research, Zurich, Switzerland; ²ETH, Zurich, Switzerland

Tumors are unique and complex ecosystems that consist of heterogeneous cancer and immune cell type populations with variable molecular profiles, different aggressiveness and proliferation potential. Understanding how heterogeneity influences tumor progression has important clinical implications for the design of effective, personalized treatment options. Towards this goal, single-cell proteomics have enabled the simultaneous quantification of dozens of proteins in millions of cells in situ, permitting a deep characterization of spatial heterogeneity at an unprecedented resolution. However, the computational analysis of the generated data is challenging, and appropriate quantitative metrics are not yet established.

We present a computational framework to quantify tumor heterogeneity in a spatial manner (Figure 1A). Starting from spatially resolved cell phenotypes, we construct a cell-cell graph that describes cell interactions and use custom heterogeneity metrics to capture the underlying tumor architecture. In a next step, we employ machine learning approaches to learn which heterogeneity scores are able to predict disease subtypes and identify the most relevant metrics driving the decision (Figure 1D). Interestingly, immune cells exhibit stronger clustering patterns in grade 3 tumors, and preliminary results indicate that immune-tumor interaction patterns are potentially prognostic for survival (Figure 1E). Our framework is extendable and applicable to other in situ approaches and has the capability to provide data-driven patient stratification or to support comparative clinical studies.

Figure 1. (A) Overview of the proposed pipeline. (B) Example tumor samples for two tumors of grade 1 (top row) and 3 (bottom row) that exhibit distinct spatial patterns and spatial entropy (C). The heterogeneity metrics used to successfully predict disease subtype (D) and linked to clinical metadata, such as grade and survival (E).

References
Implementation of a Clinical Decision Support System for Automated Fall-Risk Identification and Referrals in Emergency Departments

Apoorva Maru, BS1, Gwen Costa Jacobsohn, PhD, MA1, Collin Engstrom, PhD1, Frank Liao, PhD2, Margaret Leaf, MS2, Joel Galang2, Gerald Pankratz, MD1, Alexis Eastman, MD1, Maureen Smith, PhD, MPH1, Manish Shah, MD, MPH1, Brian Patterson, MD, MPH1

1. University of Wisconsin-Madison, Madison, WI, USA 2. UW Health, Madison, WI, USA

Introduction
While almost 3 million older adults are seen for fall-related injuries in United States emergency departments (EDs) each year1, EDs have played little role in falls prevention efforts. Existing interventions for fall risk screening in the ED have not been widely adopted in practice due to the time and resources necessary to perform in-person screening2,3. Our interdisciplinary team designed an automated EHR-based clinical decision support (CDS) tool to identify and refer older ED patients at high risk of future falls with minimal staff burden. Post-implementation evaluation and user feedback suggested that limitations in the EHR interface prevented ED providers from seeing the automated fall risk alert. Therefore, we evaluated the effect of implementing a forced interaction with the alert as a hard stop.

Methods
Figure 1 depicts the automated screening and referral intervention: a previously derived algorithmic screen automatically identifies high-risk patients, for whom a CDS alert prompts ED Physicians to refer to a fall risk reduction clinic. The workflow was designed with input from clinical stakeholders, patients, and IT staff using human factors principles including usability testing with a heuristic evaluation and cognitive walkthroughs. The intervention was piloted in an academic emergency department with over 60,000 yearly visits, beginning 7/8/2020. Initial results, while promising, showed a relatively low rate of physician acceptance of the CDS guidance, despite educational initiatives and academic detailing. Beginning on 11/16/2020, the CDS was altered to force physicians to either accept or explicitly reject a referral through a hard stop. Data collection began on 7/8/2020 and ran through 3/1/2021.

Results
During the overall study period, 3887 older adult patients were discharged from the ED. Of these, 392 patients were identified as high risk by the automated algorithm (10% 95% CI [9% - 11%]). Before the implementation of the hard stop, 190 patients were flagged and 46 were referred (24% [18% - 30%]). After the implementation of the hard stop, 202 patients were flagged and 77 were referred (38% [31% - 45%]). This difference was significant in a two-tailed test of independent proportions, p = 0.002. We collected qualitative data from 10 physician interviews analyzed via content analysis. Interview data suggested no increase in provider burden with the hard stop.

Conclusion
Despite a rigorous design process and user education campaign, as well as survey data suggesting high acceptance, intervention uptake was initially low when providers were able to ignore the CDS tool. The design team, despite initial hesitancy, added a hard stop to promote interaction with the tool. The addition of the hard stop resulted in significant increase in utilization (p=0.002) across all ED provider groups. Qualitative data indicated maintenance of high acceptability among users.

References

Figure 1. Automated Screening and Referral Process
There is Variability in the Accuracy of Diagnosis Codes Used to Identify COVID-19 Patients

Jayson S. Marwaha, MD1,2; Elizabeth Eldridge, MPH3; Sahr Syed, BS3; Perry Mar, PhD3; Philip Ballentine, BS3; Sadiqa Mahmood, DDS, MPH3; Denis Agniel, PhD2; Nathan Palmer, PhD2; Gabriel A. Brat, MD, MPH1,2

1Department of Surgery, Beth Israel Deaconess Medical Center, Boston, MA; 2Department of Biomedical Informatics, Harvard Medical School; 3Health Catalyst, Salt Lake City, UT

Introduction: On April 1, 2020, a new diagnosis code for COVID-19 was introduced in ICD-10-CM (U07.1: “COVID-19, virus identified”).1 Recent research has found that this diagnosis code has excellent accuracy in identifying hospitalized COVID-19 patients from administrative claims data. From April to May 2020, the sensitivity and specificity of U07.1 were found to be 98% and 99%, respectively.2 While these metrics from the early months of the pandemic are encouraging, the accuracy of this diagnosis code in identifying COVID-19 patients in other types of non-inpatient healthcare interactions remains unknown. As the number of journal articles using this identifier to define their cohort grows, it is important to understand whether U07.1 reliably captures COVID-19 patients in all settings. In this study, we examined the overall sensitivity and specificity of U07.1 across all types of patient interactions with the healthcare system or electronic health record (EHR). By leveraging a unique national database, we found that a large fraction of COVID-19 patients may not be identified by U07.1, despite being clinically affected by the disease and having interactions with the healthcare system. This may affect the integrity of retrospective analyses on COVID-19 patients.

Methods: Our analysis used de-identified data on 1.6 million patients in the Health Catalyst Touchstone registry, representing 17 health systems across 24 states. Patients from January 1 to July 21, 2020 with a positive SARS-CoV-2 polymerase chain reaction (PCR) test or the ICD-10-CM code U07.1 were included.

Results: 1,564,861 patients were included. Hospitalized patients accounted for 22.1% (n=345,676) of our cohort, and non-hospitalized patients accounted for 77.9% (n=1,219,185). Among hospitalized COVID-19 patients (n=345,676), the sensitivity and specificity of U07.1 were 95.6% and 97.9%, respectively. The sensitivity and specificity from April-June 2020 was 92.8% and 99.1%, respectively; from July-September they were 95.8% and 98.9%; and from October-December they were 93.6% and 98.3%, respectively. These findings are consistent with prior studies on the accuracy of U07.1 among inpatients. Importantly, these findings also show that the accuracy of U07.1 has remained steady over time. However, we found that the accuracy of U07.1 deteriorated significantly when non-inpatient interactions were examined. When we included patients with any documented health system interaction (ex. outpatient office visits or telehealth visits), the overall sensitivity and specificity for all patients (n=1,564,861) were significantly lower at only 54.1% and 98.6%, respectively.

Discussion: In this study, we found that the accuracy of U07.1 in identifying PCR-positive COVID-19 patients varies significantly by type of patient interaction. While U07.1 is a reliable identifier among inpatients, it may miss a large fraction of COVID-19 patients in other healthcare settings. Retrospective analyses that use U07.1 codes to identify COVID-19 patients for policy, surveillance, or academic purposes need to account for the significant number of patients not captured by this code. Sample cohorts may fail to analyze an important group of patients and must consider this limitation to ensure appropriate capture of patients.

References

Analyzing COVID-19 Case Report Forms Data from the All of Us Program

Craig S. Mayer, MS¹, Nick Williams, Ph.D¹, Vojtech Huser MD, Ph.D¹
¹Lister Hill National Center for Biomedical Communication, National Library of Medicine, NIH Bethesda, MD

Introduction

The All of Us (AoU) program developed a survey to assess the impact of COVID-19 on its participants.¹ The program began collecting data for the ‘COVID-19 Participant Experience’ (COPE) survey in May of 2020 with the intention for each participant to complete the survey multiple times. In this study we describe how survey data are represented in the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM), how data is redacted in the ‘registered tier (level 2)’ access level of the AoU workbench and showcase selected survey research insights.

Materials and Methods

First, using the OMOP documentation and provided relationship dictionary we identified which COPE survey questions (Data Elements [DEs]) were redacted in the registered tier (level 2) of access. Second, we analyzed the format of the survey in the OMOP CDM to identify structurally how different questions and answers are adopted to the CDM including different data types, questions with multiple answers, and related DEs. Third, we analyzed the results of the COPE survey for a cohort of possible COVID-19 patients, using the 2020 fourth quarter data release.

Results and Discussion

Redaction: The COPE survey consisted of 182 DEs from 16 topics. Of the 182 DEs, 104 (57.1%) were accessible. 78 were redacted, including five complete survey sections for COVID-19 Treatment, COVID-19 Testing, Mood, Stress, and Social Support. All EHR DEs relating to COVID-19 testing and treatment were also redacted.

Data type and structure: Based on data type, 87 (83.7% of the 104 accessible DEs) are categorical, 13 (12.5%) are numerical and 4 (3.8%) are free text stings. Structurally, for questions that can have multiple answers selected, such as ‘In the past month, has the COVID-19 outbreak affected you? Please select all that apply’, there is a data row for every answer provided. If a participant selects both ‘worked remotely’ and ‘worked reduced hours’, the responses will appear on two separate rows for the one response. We also saw two types of related DEs, grouped (DEs that cannot stand alone) and branched (DEs which are dependent on the response of another DE), where the only connection between the DEs is seen in the relationship dictionary. An example of grouped DEs were ‘How long has it been since you last smoked’ and ‘Enter the number of years’, whereas an example of a branched DE is ‘Do you know someone who has died from COVID-19?’, which triggers ‘Who do you know who has died?’.

Cohort: As of the 2020 fourth quarter data release, 62,664 people have completed the COPE survey, compared to 315,297 (19.9%) who have enrolled in the AoU program so far. To demonstrate the COPE survey data research potential, we considered a cohort of 3,700 possible COVID-19 patients based on the answering of ‘Yes’ to the unredacted ‘Do you think you have had COVID-19?’. The most common reported symptoms from this cohort were ‘Unusual fatigue’ (73.5% of participants), ‘Cough’ (64.1%), and ‘Headache’ (62.6%). The biggest difference in responses from the COVID-19 cohort was in answering ‘Have you EVER been near someone that you know, or suspect, had COVID-19?’. The COVID cohort answered ‘Yes, suspected COVID-19’ (16.4% increase) and ‘Yes, known COVID-19’ (15.5% increase) far more often. Additional results are available in the project repository at https://github.com/lhncbc/CRI/tree/master/AoU/COPE.

AoU R package: We extended our previously developed R package (r4au) for frequently needed AoU workbench analytical tasks with additional functions (getCaseReportForms and getRegisteredDataDictionary) that, respectively, generate the dictionary and label the DEs with the topic and module (form) that they come from.

Conclusion: We concluded that the COPE survey data was able to be fully stored within the OMOP OBSERVATION table and the results analyzable in the research workbench. Both are significant demonstrations for other studies considering the OMOP model and adopting similar conventions used by AoU. AoU’s ability to re-contact participants also represents a powerful platform for future research as proven by COPE survey follow-ups.

References

Barriers to Seeking Treatment and Attitudes Towards Using Telehealth to Support Management of Depression Among Black Women

Terika McCall, PhD, MPH, MBA
1Center for Medical Informatics, Yale School of Medicine, New Haven, CT

Introduction
Approximately 1 in 5 Black women in the United States experienced mental illness in the last year (1). Furthermore, depressive disorders are the most common mental health conditions among Black women (2). Barriers to seeking treatment contribute to Black women using mental health services at less than half the rate of their White counterparts (1). Findings from numerous studies have shown that the use of telehealth interventions (eg, mobile app) can increase access to services and resources, and are effective in helping participants reduce depressive symptoms (3,4). This study aimed to identify barriers to seeking treatment and gain insight into which modalities Black women deem acceptable for use to communicate with a professional to receive support with managing depression.

Methods
The web-based survey was launched in October 2019 and closed January 2020. Black/African American women (≥ 18 years old) were eligible to participate. Participants were recruited through convenience sampling (eg, email sent via listservs). The 80-item survey included questions about barriers to seeking mental health care and the acceptability of using mobile technology to communicate with a professional to receive support for depression. Fisher’s exact test was used to determine whether an association exists between the responses to questions about comfortability with using each modality to communicate with a professional to receive help for managing depression and age group and education level, respectively. Statistical significance was determined at the 2-sided p<.05 level for all tests.

Results
Findings from the study (N=395) showed that Black women had favorable views toward seeking mental health care, however approximately 40% of respondents indicated that during the past 12 months there was a time when they needed mental health treatment or counseling but did not receive care. Identified barriers to receiving care included cost, not knowing where to access services, lack of time, and stigma. Overall, respondents were more comfortable with using voice call (69.9%) or video call (64.3%) when compared with text messaging (45.1%) or use of a mobile app (45.1%) to communicate with a professional for support with managing depression. Younger Black women (< 50 years old) were significantly more likely to endorse use of text, voice call, mobile app, or video call to communicate with a professional to receive support for managing depression (all p<.05). Furthermore, those with less than a Bachelor’s degree were significantly more likely to endorse the use of text messaging (p=.026). No significant associations were found between education and other modalities (all p>.05).

Conclusion
The findings of this study may be used to inform design of telehealth interventions for supporting management of depression among Black women. Incorporating their preferences for telehealth modalities to communicate with a professional to receive support may improve treatment outcomes.

References
Addressing the Social Determinants of Health and Resource Utilization Among Ambulatory Oncology Patients

Nadine J McCleary MD MPH,1,2 Ellana Haakenstad MPH,1 Sunyi Zhang MS,1 Jessica LF Cleveland MSOR,1 Michael Manni,1 Deb Schrag MD MPH,1,2 Michael Hassett MD MPH1,2

1Dana-Farber Cancer Institute, Boston, MA; 2Harvard Medical School, Boston, MA

BACKGROUND: Unplanned emergency department (ED) and hospitalizations (hospit) disrupt care for cancer patients (pt) and may be driven by SDoH needs. To develop an intervention to address SDoH in oncology, we sought to demonstrate the systematic collection of SDoH at an academic medical center. We investigate if ED/hospit rates differ based on pt demographics and SDoH needs regardless of cancer diagnosisstage.

METHODS: We conducted a cross-sectional analysis of the SDoH among new patient oncology consults from 5/30/2015 – 4/30/2020 at Dana-Farber Cancer Institute (DFCI), Boston, MA. Patients were invited to complete a New Pt Intake Questionnaire (NPIQ) eliciting SDoH including: living alone, financial distress (“How difficult is it for you to meet monthly payments on your bills”), health literacy (“How confident are you filling out medical forms”) and health numeracy (“How confident are you in understanding medical statistics”). Sociodemographics were self-reported: race/ethnicity (White, Black, Latinx, American Indian or Alaska Native (AIAN), Native Hawaiian or Pacific Islander (NHPI)), limited English-proficiency (LEP), and age (<40, 40-69, ≥70 years). Multi-ethnic was not included as a race/ethnicity option. The electronic health record was queried to identify all ED visits and hospit occurring within 30 days of the initial oncology visit. The association between SDoH and ED/hospit subsequent to the oncology visit were determined with χ² tests.

RESULTS: Of the 145,233 new consults evaluated between 5/30/2015 – 4/30/2020, 14.1% identified as Black, AIAN or NHPI; 0.7% Latinx; 5.5% LEP; and 25.8% older adults (age ≥70 years). Of these, 38,372 (42.8%) were evaluated in an MGB ED or hospital and 7,089 (0.7%) identified at least one SDoH need. ED significantly differed by race (37.2% Black vs. 10.7% white), ethnicity (59.1% vs. 17.1 % non-Latinx), LEP (33.9% vs. 17.3% non-LEP), and older adults (18.8% vs. 16.7% age 40-69 vs. 18.4 age <40). Except for health literacy, SDoH were also significantly associated with ED regardless of demographics. Except for living alone, similar findings were found for hospit.

CONCLUSION: Identifying SDoH in oncology is imperative for interventions that reduce disruptions to care and improve outcomes, particularly for pt subsets. Approximately 22.0% of pts responded to at least one SDoH question on the NPIQ. We are leading a study that will utilize a targeted, systematic SDoH needs screening tool for all pts allowing for a targeted intervention to document and intervene on SDoH needs in oncology.

Table 1: Resource utilization by social determinants of health among ambulatory oncology patients

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>ED visit (%)</th>
<th>p-value</th>
<th>Hospitalization (%)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18-39 years</td>
<td>18.4</td>
<td>&lt;0.001</td>
<td>34.5</td>
<td></td>
</tr>
<tr>
<td>40-69 years</td>
<td>16.7</td>
<td>&lt;0.001</td>
<td>37.2</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>≥70 years</td>
<td>18.8</td>
<td></td>
<td>40.3</td>
<td></td>
</tr>
<tr>
<td>Race</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Black</td>
<td>37.2</td>
<td>&lt;0.001</td>
<td>48.8</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>AIAN</td>
<td>22.6</td>
<td></td>
<td>35.7</td>
<td></td>
</tr>
<tr>
<td>NHPI</td>
<td>26.9</td>
<td></td>
<td>38.5</td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>10.7</td>
<td></td>
<td>30.6</td>
<td></td>
</tr>
<tr>
<td>Asian</td>
<td>18.8</td>
<td></td>
<td>38.6</td>
<td></td>
</tr>
<tr>
<td>Ethnicity</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Latinx</td>
<td>59.1</td>
<td>&lt;0.001</td>
<td>51.8</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Non-Latinx</td>
<td>17.1</td>
<td></td>
<td>37.4</td>
<td></td>
</tr>
<tr>
<td>Language proficiency</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LEP</td>
<td>33.9</td>
<td>&lt;0.001</td>
<td>48.3</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>EP</td>
<td>17.3</td>
<td></td>
<td>38.3</td>
<td></td>
</tr>
<tr>
<td>Live alone</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>20.3</td>
<td>&lt;0.001</td>
<td>41.9</td>
<td>0.085</td>
</tr>
<tr>
<td>No</td>
<td>17.4</td>
<td></td>
<td>40.1</td>
<td></td>
</tr>
<tr>
<td>Health literacy</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>18.1</td>
<td>0.712</td>
<td>43.6</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>No</td>
<td>17.8</td>
<td></td>
<td>39.6</td>
<td></td>
</tr>
<tr>
<td>Health numeracy</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>16.5</td>
<td>0.004</td>
<td>41.9</td>
<td>0.005</td>
</tr>
<tr>
<td>No</td>
<td>18.3</td>
<td></td>
<td>39.6</td>
<td></td>
</tr>
<tr>
<td>Financial distress</td>
<td></td>
<td>&lt;0.001</td>
<td></td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Yes</td>
<td>20.7</td>
<td></td>
<td>44.8</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>No</td>
<td>16.9</td>
<td></td>
<td>40.1</td>
<td></td>
</tr>
</tbody>
</table>
Electronic Health Record-Based Risk Stratification for Recurrence of Kidney Stones: A Feasibility Implementation

Allison B. McCoy, PhD, Ryan S. Hsi, MD
Vanderbilt University Medical Center, Nashville, TN

Introduction

Symptomatic kidney stone events occur in 1 in 11 Americans, and repeated events occur in up to 50% of individuals within ten years. These recurrence episodes contribute to adverse health outcomes, detriments in health-related quality of life, and increased costs of care. Since the risk of future stone events varies for each individual, it is necessary to accurately risk stratify patients to tailor the intensity of preventative treatment to reduce stone recurrence. The current paradigm of kidney stone care relies on provider-level experience for risk stratification. Clinician-predicted stone recurrence risk varies widely among providers, and these estimates significantly differ from risk predicted by a validated nomogram (Recurrence of Kidney Stone [ROKS]). Meanwhile, adoption of nomograms in clinical practice has not occurred due to lack of knowledge on their use and high burdens of work to use nomograms that are implemented in external applications, despite clinician willingness to use such tools. We present a feasibility implementation of the ROKS nomogram into the electronic health record (EHR), which we term EHR-ROKS, to improve provider risk stratification and delivery of appropriate preventive treatment for patients experiencing kidney stone events.

System Description

In our feasibility implementation of EHR-ROKS into our Epic EHR, the nomogram is accessed from within the progress note using a text shortcut (i.e., SmartPhrase) that inserts a hyperlink into the note, as well as the predicted risk of kidney stone recurrence at 5 and 10 years, once calculated. Clicking the hyperlink will open the ROKS questionnaire in a flowsheet activity that allows clinicians to answer questions about the patient and prior stone episode and automatically calculate and update the predicted kidney stone risk in the note.

The ROKS nomogram sums total points based on 10 questions (or 14 if imaging was performed at the last stone event), then calculates risk as $1 - \alpha e^{-\alpha_5(1.84099 + \text{points} \times 0.019)}$ with defined values for $\alpha$ by stone episode and number of years since the last stone episode. The questions include age, BMI, sex, family history of kidney stones, presence of an incidental prior stone, suspected prior kidney stone, pregnant during last stone event, stone composition, number and size of prior stones, and stone location.

Within Epic, calculations for EHR-ROKS are performed using a combination of 18 logic rules that return numeric values. Two rules impute age and BMI, and one rule determines whether the patient is pregnant. Another rule determines the total number of points based on the three prior rules and the manually entered flowsheet values. To determine the $\alpha$, five rules return the $\alpha$ for 5-year risk, and five rules return the $\alpha$ for 10-year risk based on the manually entered episode number and number of years since the last stone episode. Two rules calculate the 5- and 10-year risk, respectively.

Conclusion

Our feasibility implementation indicates that the ROKS nomogram can be integrated into our commercial EHR into existing workflows, improving the likelihood of adoption in future work to improve provider risk stratification and delivery of appropriate preventive treatment and ultimately reduce kidney stone recurrence. Similar nomograms could be implemented using this approach in other clinical scenarios.

References

Approaches to Clinical Decision Support Alert Evaluation and Optimization

Allison B. McCoy, PhD, Adam Wright, PhD
Vanderbilt University Medical Center, Nashville, TN

Introduction
Despite extensive research into implementing and evaluating clinical decision support (CDS) alerts within electronic health records (EHRs), many continue to deliver alerts that are sub-optimal; most alerts are irrelevant and repetitive, creating additional work for clinicians and leading to considerable EHR-related information burden. This burden contributes to clinician burnout, which occurs as a reaction to long-term occupational stress exposure, and can adversely affect patient safety and care quality. Optimization of CDS and other EHR capabilities can reduce EHR-related burden, but current processes for developing, maintaining, and delivering such complex CDS alerts can be labor- and cost-intensive. We sought to review existing approaches for identifying opportunities to optimize alerts.

Methods
Through review of the literature and author experience, we identified previously applied approaches to evaluating and optimizing CDS alerts. For each approach, we determined the optimization opportunities that could be identified, the technical or build effort required to implement the approach, and the personnel effort required.

Results
One of the most frequently described approaches to automatically identify sub-optimal alerts in prior literature has involved individual or committees reviewing EHR utilization logs or dashboards, as these are very effective at identifying individual alerts with high override rates, as well as alert trends across departments or provider type. However, efforts required to develop reports or dashboards can be high, and personnel effort required to review the findings are substantial. Commercial dashboard products and analytics tools directly in the EHR can alleviate some of the technical burden, but personnel effort remains high. An initiative to optimize CDS alerts through detailed review called “Clickbusters” encouraged participants who were physicians, pharmacists, nurses, or informaticists with training or interest in CDS implementation to follow ten steps to improve CDS alerts. Additional efforts have focused on identifying opportunities for improvement by obtaining feedback from clinicians receiving the alerts, including sentiment analysis of comments entered by clinicians when overriding alerts and feedback links within displayed alerts allow clinicians to easily convey positive, neutral, or negative feedback to CDS team members about well- or poorly-performing alerts. Like EHR utilization log and dashboard review approaches, these approaches successfully identify individual alerts with optimization opportunities, but they can also identify alerts with potentially sub-optimal design, or those that are no longer functioning as designed. These approaches also have a slightly lower barrier with build and personnel effort, as they utilize a crowdsourcing approach to gaining the information and only require personnel to monitor feedback given. However, these approaches are limited in that the only evaluate alerts that clinicians see and to which they respond. Anomaly detection to identify alerts that are no longer performing as intended can successfully identify all alerts that are no longer functioning as designed, but with less personnel effort due to the machine learning approach; however, it does not identify specific opportunities for improving alerts that were sub-optimally designed at their onset. A recent tool released by Epic (“BPA TuneUp”) evaluates single characteristics (e.g., visit diagnosis, provider type) for each implemented alert and determines the number of actions that could be taken and lost by excluding that characteristic. This approach has the potential to identify all alerts with sub-optimal design with low technical and personnel effort; however, this tool remains limited in the types of features evaluated, it cannot account for multiple features or scenarios, and it is proprietary and not generalizable to other EHRs or EHR components.

Discussion
Diverse approaches for evaluating and optimizing CDS alerts exist, each with distinct advantages and disadvantages as to the types of improvement opportunities that can be identified and the requirements for adopting each approach. However, these approaches most often require significant personnel effort, and they fail to comprehensively capture potential scenarios for improvement across all alerts and all CDS implementations. Future research is needed to develop effective, efficient, generalizable methods that can better identify opportunities to optimize CDS alerts and implement these improvements to reduce EHR-related information burden and clinician burnout.
Rolling up the Sleeve: Equitable, Efficient, and Safe COVID-19 Mass-Immunization for Academic Medical Center Employees

Samuel McDonald, MD, MS1, Mujeeb A Basit, MD, MMSc1, Seth Toomay, MD1, Christopher McLarty, D.N.P., APRN, ACNP-BC1, Susan Hernandez, D.N.P., MBA, RN1, Chris Rubio, MBA1, Bruce J. Brown, Dr.P.H.1, Mark Rauschuber, MHA1, Ki Lai, MS1, Sameh N Saleh1, DuWayne L Willett, MD, MS1, Christoph U. Lehmann, MD1, Richard J Medford, MD, FRCP(C)1

1University of Texas Southwestern Medical Center, Dallas, Texas, USA

Introduction: Since the emergence of the SARS-CoV-2 virus and the ensuing global pandemic, governments of affected countries have focused on decelerating the spread of COVID-19 infection until the deployment of effective vaccines or treatments with shifting attention of US governments and hospitals to vaccine distribution, delivery, and acceptance. It is our hypothesis that a Prioritization/Invitation/Scheduling framework would 1) provide an efficient, infrastructure to optimize speed and efficiency of vaccine delivery while minimizing risk of infection during the immunization process and 2) provide insight into vaccine acceptance among health care workers.

Methods: This was an observational study with data collection occurring during the early phases of COVID-19 vaccination availability. We implemented a Prioritization/Invitation/Scheduling system that allocated vaccines based on exposure risk and stepwise invited employees to schedule the vaccine administration through the health system’s patient portal. Employees were categorized into six groups: providers, nurses, ancillary healthcare workers, facility & food services, research staff, and support staff. We measured the period from invitation to scheduling the vaccination appointment through the patient portal as a proxy measure for the level of interest in receiving the COVID-19 vaccination and stratified data by phase and job category. To assess operational efficiency, we evaluated the time from throughput and the number of vaccinations completed per hour over the first week of vaccinations.

Results: We allocated vaccines to 13,753 employees. Ultimately, 10,662 employees had or added portal access and received a vaccination invitation. During the allocation phase, we categorized 2,473 employees (18%) into the highest priority (Phase 1a.1). The majority were providers and nurses, but ancillary clinicians and administrative staff (e.g. unit clerks) were also included. While the Phase 1a.2 cohort had a similar distribution, the Phase 1a.3 cohort had 649 facility and food workers and 996 administrative staff. Out of those employees receiving an invitation, 6,483 (61%) scheduled an appointment and 6,251 (59%) employees were immunized in the first seven days of availability. Employees varied significantly by type. About 66% of invited providers were vaccinated in the first seven days, whereas only 41% of facility/food service employees received the first dose of the vaccine (p<0.001). Clinical research staff and providers had the highest uptake (68% vs. 66%, p = 0.97). Providers were the fastest group in scheduling their vaccines (92 minutes) with the longest to schedule were facility/food service employees (588 minutes). 6,251 vaccines were administered during the study period with average of 60 vaccinations per hour. Throughput was found to have average time of 5.6 minutes (IQR 3.9 – 8.3) with slightly longer wait times the first (6.8 minutes). No difference was found in wait times across any group.

Discussion: The historically short interval from vaccine development to authorization, the politicization of the pandemic and pandemic measures, the paucity of accurate data throughout the pandemic as well as historic distrust of the medical establishment in minority communities are reflected in preliminary surveys that suggest many Americans are reluctant to “roll up their sleeves” to be vaccinated. Despite this, we found an interest among healthcare workers to receive the vaccination at the earliest possible opportunity. Even though invitations were delayed, some employees realized that scheduling was available to them in the mobile patient portal and word spread quickly to others. We noticed a stark difference in the willingness to be vaccinated by the different employee groups. It is possible employees in the later phases may have received a vaccine invitation from the affiliated organizations. The dwell time for employees at the vaccination site was very short and improved to under 5 minutes (excluding observation) as the week progressed. This minimized the risk for infection during the vaccination.

Conclusion: We offer early data on vaccine prioritization, administration, and uptake in our institution with the intent of offering insight and a potential model for other health systems to adopt for not only employee vaccination but also one that is scalable to vaccinate the general populous in an efficient and safe manner. We hope that our findings are not a unique response within our health system and are instead a marker for not only other health systems but also the rest of the country.
Design and Feasibility of ACTIVATE, a Digital Health Platform for Community Health Centers in Response to COVID-19

Scott P. McGrath, PhD1; Cynthia G. Matsumoto, PhD, RN2; David Lindeman·PhD1;
Katherine K. Kim, PhD, MPH, MBA, FAMIA3,4

1University of California Berkeley, CITRIS and the Banatao Institute, Berkeley, CA, USA; 2 University of California Davis Health, Sacramento, CA, USA; 3University of California Davis, Betty Irene Moore School of Nursing, Sacramento, CA, USA; 4University of California Davis, School of Medicine, Department of Public Health Sciences/Division of Health Informatics, Sacramento, CA, USA

Introduction
Rural, low-income agricultural workers in California, many of whom are undocumented, are especially vulnerable to ill health, including COVID-19. Rural populations suffer significant health disparities arising from a variety of circumstances, notably including multiple adverse social determinants of health and inadequate access to high-quality healthcare. The recent COVID-19 pandemic has exacerbated health inequities and digital disparities in areas such as California’s rural and agricultural region called the Central Valley. ACTIVATE: Accountability, Coordination, and Telehealth in the Valley to Achieve Transformation and Equity was conceptualized and designed to bridge these specific disparities and to accomplish two key aims: co-design a digital health solution for community health centers and assess the feasibility of implementation in one center in the Central Valley.

Methods
Two stages will be focused on in this poster: a co-design stage and a feasibility pilot stage. Feasibility was determined if at least 50% of participants fulfilled an established data upload frequency within three weeks.

Results
Co-design: The co-designers informed delivery of educational material, remote patient monitoring equipment set up, and a technology assessment tool. They also tested technology. Co-designers also helped refine a technology readiness assessment tool sharing their concerns and practical considerations that might affect patient readiness.

Feasibility pilot: 12 patients recruited for the pilot (out of 15 contacted). The ages range from 38-64. 11 are female and 9 are monolingual Spanish speakers with the remaining 3 being English speakers. 3 participants have diabetes, 3 have hypertension, and 6 have both. For Internet connectivity, 11 patients have smartphones, 3 of the 11 with phones did not have a data plan and only 5 of 12 have broadband access at home. In 12 weeks of the feasibility pilot period, 1,046 unique vital measures were transmitted including 691 blood glucose values and 355 blood pressure values. 3 of 9 the patients with diabetes transmitted 18 measures in their first week and 3 of 9 patients with hypertension transmitted 14 measures in the first week. By the third week, 4 of 9 of the diabetes patients had submitted 18 entries and 5 of 9 of the hypertension patients had 14 entries. Thus, the feasibility pilot cleared the threshold of 50% of enrolled participants by week 3.

Discussion and Conclusion
Recruitment into the program in for the feasibility pilot was very high. The large number of measures transmitted and high level of patient engagement with health coaches bodes well for future expansion. A key goal of ACTIVATE is to demonstrate the program is scalable and repeatable in other underserved communities. COVID-19 only heightened existing socioeconomic inequalities and the digital divide in underserved populations. Individuals who already had difficulty visiting healthcare providers in person, were faced with additional hurdles when clinics and physicians shifted to telemedicine and remote care. The safety and convenience of telemedicine only benefited those who had the technology and connectivity to participate. ACTIVATE targets these very inequalities with direct patient engagement, hardware and Internet access solutions, and increased health literacy to improve outcomes all through a digital health based solution. Experience in the pilot indicates that delivering on ACTIVATE’s key goals are feasible. ACTIVATE is currently being expanding in the Central Valley and plans include development of tools for dissemination of the model elsewhere.
Natural Language Processing and COVID-19 Predictive Analytics to Enable and Optimize SARS-CoV-2 Pooled Testing

Stephane Meystre, MD, PhD, Paul M. Heider, PhD, Jihad Obeid, MD, Alexander Alekseyenko, PhD, James Madory, DO

Medical University of South Carolina, Charleston, SC

Introduction: As part of the COVID-19 pandemic response at the Medical University of South Carolina (MUSC), a telehealth system was implemented as the preferred option for patients interested in COVID-19 diagnostic testing. Patients would start a virtual visit, indicate their symptoms, COVID-19 exposure and travel history, and brief medical history. After completion of this virtual visit, the telehealth system exports a summary text note generated from the information entered by the patient. To enable multiple efforts supporting the care and management of patients diagnosed with COVID-19, structured and coded data was required and a new dedicated natural language processing (NLP) application (COVID NLP tool) was therefore developed, evaluated and deployed in March-April 2020. This COVID NLP tool was subsequently improved (as described below) and processes text notes stored in a new datamart focused on COVID-19 (COVID-19 datamart). Early success with using information from the telehealth system to predict positive SARS-CoV-2 results encouraged further efforts to enhance the accuracy of these predictions and enable applications supporting patient care such as a novel data-driven COVID-19 symptom checker giving patients testing advice according to their predicted test result. Further experiments with NLP and predictive analytics for SARS-CoV-2 test results prediction with various behaviors and optimization experiments followed along with applications to improve the capacity and efficiency of SARS-CoV-2 diagnostic testing by enabling pooled testing in conditions typically preventing it, as described below.

Methods: The COVID-NLP tool combines components we had developed in past efforts and could reuse with components from other local current research (rule-based or deep learning-based, not retrained) and a few new custom components (i.e., for data import/export, section detection, and key demographics). The deep learning components are based on Bi-LSTM sequence labeling models using word tokens as input and had been trained on publicly-available annotated corpora. Vector representations of words were constructed using fastText embeddings pre-trained with clinical texts (MIMIC-III). To evaluate the COVID NLP tool accuracy, a small set of 15 text notes was randomly selected and manually annotated by domain experts (two independently and a third adjudicating differences).

For the development, optimization and evaluation of predictive analytics, we included all patients using the telehealth system for virtual care visits at MUSC in April-June 2020 who had a SARS-CoV-2 diagnostic test within 14 days after the virtual visit (14,055 patients, 1,101 testing positive and 12,954 negative). Information extracted by the COVID-NLP tool was then used for predicting the SARS-CoV-2 test result. To select algorithms for prediction, shallow and deep learning algorithms were compared. The initial features used for prediction included age, a selection of 23 symptoms and deep learning components are based on Bi-LSTM sequence labeling models using word tokens as input and had been trained on publicly-available annotated corpora. Vector representations of words were constructed using fastText embeddings pre-trained with clinical texts (MIMIC-III).

Results: The COVID-NLP tool accuracy results are presented in Table 1. For predictive analytics, we focused on SVM and logistic regression models in the later stages of this work. The best logistic regression evaluated with our telehealth data had a recall of 0.9514, specificity of 0.1244, and negative predictive value (NPV) of 0.9286.

<table>
<thead>
<tr>
<th>Table 1: COVID NLP tool prototype accuracy results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Environmental Risk Factors</td>
</tr>
<tr>
<td>Demographics</td>
</tr>
<tr>
<td>Medical Risk Factor</td>
</tr>
<tr>
<td>Medications</td>
</tr>
<tr>
<td>Micro-average</td>
</tr>
<tr>
<td>Macro-average</td>
</tr>
</tbody>
</table>

References

Transparency with Safeguards: Suppressing Notes from a Patient Portal

Chelsea L. Michael, MS\textsuperscript{1}, Gilad J. Kuperman, MD, PhD\textsuperscript{1}
\textsuperscript{1}Memorial Sloan Kettering Cancer Center, New York, NY

Introduction

The Information Blocking Rule (IBR) of the Cures Act has increased attention to how healthcare organizations provide electronic health information to patients\textsuperscript{1}. The IBR coincided with our phased project to add clinic notes to the patient portal (portal) at Memorial Sloan Kettering Cancer Center (MSK). For our project, the experts at OpenNotes urged us to create a feature so providers could prevent a specific note from appearing in the portal (if it would appear by default). While it was a best practice, OpenNotes had limited information about other organizations’ features and usage. They recognized a gap in the literature. Therefore, we aimed to implement and evaluate a feature to allow a clinician, when warranted, to prevent a note from appearing in the portal, within our project to post notes to the portal by default.

Methods

MSK is a dedicated cancer center. Our electronic health record (EHR) is Allscripts Sunrise and our portal is custom. To create the capability to prevent notes from appearing in the portal, we gathered version 1 (v1) requirements with input from the project team, Health Information Management, the MSK clinical documentation team and the EHR optimization committee. We created and piloted the v1 feature in 3 services for 5 months, starting in late 2020. We trained users through an email at go-live and posted additional resources to our intranet. We evaluated the feature through a discussion with Ethics, Compliance/Legal and users; user feedback; and overall use. Finally, we created requirements for version 2 (v2) in early 2021.

Results

Our v1 requirements were: (1) allow the clinician to suppress a specific note from appearing in the portal, based on risk of harm to the patient (the feature would not impact release under HIPAA); and (2) the final note in the legal record would not indicate the note was suppressed from the portal. Our EHR did not have a built-in feature to enable suppression but we were able to implement a check box the clinician could use at the time of note creation that would suppress the note from appearing in the portal. The check box would not appear on the note output. We implemented the check box with a brief description and guidance to clinicians to only use based on risk of harm to the patient.

The pilot and evaluation led to 5 changes for v2. (1) Users wanted to know if a patient had a prior suppressed note. Therefore, we will create a way for users to check for prior suppressed notes in “writing” mode or in a separate ad hoc search application. (2) Ethics and users felt patients should be involved in the decision to suppress a note, and 3 pilot patients asked for notes to be suppressed. We will update the field explanation and training to direct users to follow patient wishes. (3) Low feature use (19 notes out of 6600+ patients with 1+ note) as expected, however half was accidental use (the provider didn’t recall using feature for 9 notes). We will create advisory text that will appear in the “writing” user interface when the checkbox is selected. (4) One service chief made a policy to consider additional referrals (e.g., Social Work, Psych) to meet patient needs when suppressing a note due to risk of harm. We will include this in the advisory text and training documents. (5) Finally, due to field properties, once suppressed, a note cannot be unsuppressed (the note cannot later be sent to portal). We noted the technical limitation in v1 training, but one user tried to reverse a suppression. We will now note the limitation in the advisory text, as well. (A note that flows to the portal can be removed through a later suppression, though.)

Conclusion

While adding clinic notes to our portal, we implemented a way to prevent a specific note from flowing to the portal based on clinician judgment of a risk of harm or patient wishes. Because our EHR did not include a feature, we used its capabilities to create one. While the IBR does not require proactively providing notes to patients, our use cases and user guidance align with IBR exceptions: risk of harm and privacy (patient’s request sub-exception)\textsuperscript{1}. At submission, we are implementing v2 and expanding to the rest of the organization.

References

Patient-Reported Visit Modality Preferences During a Pandemic

Andrea Millman, MA¹, Jie Huang, PhD¹, Yazan Nagi, MSc¹, Judy Shan, BS¹, Anjali Gopalan, MD, MS¹, Emilie Muelly, MD², Catherine Lee, PhD¹, Loretta Hsueh, PhD¹, Ilana Graetz, PhD³, Nayana Bhaskar, BS¹, Mary E. Reed, DrPH¹
¹Kaiser Permanente Division of Research, Oakland, CA; ²Kaiser Permanente Santa Clara Homestead Medical Center, Santa Clara, CA; ³Emory University, Atlanta, GA

Introduction

Telemedicine visits may offer patients convenient access to care and minimize exposure to infectious diseases. It may be unsurprising then that the COVID-19 pandemic prompted a shift toward telemedicine. However, it is unclear whether increased telemedicine use will continue post-pandemic or how experience with a telemedicine visit will impact patient visit modality preference. We surveyed patients about their visit modality choice before, and during the ongoing COVID-19 pandemic and the care settings they anticipate wanting once the pandemic ends.

Methods

This cross-sectional patient survey study was conducted within Kaiser Permanente of Northern California (KPNC). During the pandemic-related study period, patients wishing to self-schedule a visit with their primary care provider using the patient portal website or mobile app were able to choose a video or telephone visit (in-person visits required provider recommendation). We surveyed 1000 adult patients (among 1855 invited participants) who had self-scheduled a telephone or video visit in the previous month between April, 2020 and February, 2021. Respondents were 49% non-white, 40% male, 20% age 18-40, 38% age 40-60 and 42% over age 60. The survey included both structured and free-text responses. We asked patients to report their first choice visit modality before, during and (hypothetically) after the COVID-19 pandemic for a health concern that could potentially be helped by all three of the visit types and to explain their visit choice at each time point. Free-text responses were qualitatively analyzed using thematic analysis to supplement findings from structured data.

Results

Participants were more likely to report a preference for telemedicine visits during the COVID-19 pandemic than before (from 13% to 46% for video; from 18% to 31% for telephone). Expected preference for video visits after the COVID-19 pandemic was also higher than before (from 13% to 23%). About 90% of respondents provided free text responses to explain their visit modality preferences. Analysis of free text responses indicate that 63% of participants either expect to prefer, or to be open to including telemedicine visits as a care-seeking option, depending on the specific health concern and their own perceived visit barriers (Table 1).

<table>
<thead>
<tr>
<th>Pattern of visit type preference</th>
<th>Example of free text response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consistently prefers in-person visits</td>
<td>“Because I cannot get a comprehensive examination through video or telephone”</td>
</tr>
<tr>
<td>Prefers in-person, but open to Telemedicine</td>
<td>“If I am just renewing an Rx or something simple, a phone visit is fine.”</td>
</tr>
<tr>
<td>Preference changed from in-person to Telemedicine</td>
<td>“Prior to COVID-19, I wasn’t unaware of the Video visit. Now that I’ve used it, I don’t want to go back to in-person visits unless it is necessary.”</td>
</tr>
<tr>
<td>Consistently prefers Telemedicine</td>
<td>“If I can get care with a video or telephone visit, then I do not have to get a babysitter or take young kids in the car with me to a medical facility.”</td>
</tr>
</tbody>
</table>

Discussion

Post COVID-19, more survey participants expect to prefer telemedicine visits than before, and the majority will incorporate the use of telemedicine for some health care concerns, suggesting pandemic-related exposure may lead to continued use post-pandemic. Although the survey was limited to patients scheduling a telemedicine visit during the pandemic period, our study includes participants who did not initially prefer telemedicine. Offering visits by telephone, video or in-person may support patients’ ability to choose the visit types most appropriate for their health concern.
Automated Clinical Shift Schedule Generation for Emergent Care Settings

Azadeh Mobasher, PhD, Naqi Khan, MD, MS, David Elop, BASc, Zach Griffin, MBA, MHA, Naveen Valluri, MS

1Microsoft, Redmond, WA, USA, 2Providence Health & Services, Renton, WA, USA

Introduction

Clinical departments traditionally use a series of organization-specific manual processes to determine their staffing needs. These processes can equate to weeks of manual effort, while yielding sub-optimal, costly staffing plans. We propose leveraging advances in operations research, throughput simulation, and scenario prediction methodologies to augment staffing decision-making processes.

Methods

Successful, flexible shift planning for clinical staff with varying experience levels must consider organizational staffing policies, such as shift lengths, inter-shift rest durations, total weekly working hours, among many other constraints. We propose a mixed-integer program and a constraint programming based column generation algorithm to achieve a flexible shift schedule, while accounting for business logic and practical requirements such as taking downtime, working unrealistic shift lengths, and activating staff on breaks, which are unaddressed in previous work. Figure 1 illustrates the general flow of the proposed algorithm.

![Figure 1. High level flow of constraint programming-based column generation algorithm](image)

Results

We contrasted our proposed optimal scheduling costs with manually derived scheduling costs for real-world data from three emergency departments.

Table 1. Retrospective analysis for a daily planning horizon scenario.

<table>
<thead>
<tr>
<th>Emergency Department</th>
<th>Manual assignment</th>
<th>Optimal assignment</th>
<th>Daily savings ($)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Department 1</td>
<td>44</td>
<td>40</td>
<td>1,202</td>
</tr>
<tr>
<td>Department 2</td>
<td>35</td>
<td>27</td>
<td>1,484</td>
</tr>
<tr>
<td>Department 3</td>
<td>22</td>
<td>20</td>
<td>147</td>
</tr>
<tr>
<td><strong>Total daily savings</strong></td>
<td><strong>2,833</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

References

Linking U.S. Administrative Claims Data and Electronic Health Records to Describe Association Between Ambulatory Care Appointment No-Show and Healthcare Resource Use and Cost

Iman Mohammadi, PhD\(^1\), Zulkarnain Pulungan, PhD\(^1\), Nathan Markward, PhD\(^1\)

\(^1\)Avalere Health, An Inovalon Company, Washington, DC

**Introduction.** Appointment no-show is a key driver of inefficiency and waste in healthcare, contributing to incomplete and/or fragmented care delivery, reduced access to preventive services, and, consequently, sub-optimal clinical outcomes\(^1\). The objective of this study is to describe the relationship between appointment no-show and critical healthcare resource utilization (HRU) and cost among patients with Bipolar Disorder I (BPD). Data. The study utilized Inovalon’s MORE\(^2\) Registry, a large multi-payor, commercial claims repository, and a nationally representative electronic health records (EHR) database. **Methods.** Enrollment and claims data were used to identify patients who were \(\geq\) 18 years of age and newly diagnosed with BPD (ICD-10 diagnosis codes: F31.xx) during 2016. Each patient’s initial BPD diagnosis date was defined as his or her index date. Patients with diagnosis of BPD 12 months prior to the index date were excluded. In turn, patients were required to have 1) one inpatient or two outpatient diagnoses of BPD; 2) \(\geq\) 12 months of continuous enrollment pre-index; 3) \(\geq\) 24 months continuous enrollment post-index; and 4) \(\geq\) 1 EHR encounter during the follow-up period. Final study sample included 3,811 BPD patients. Using EHR appointment data, appointments were grouped into ‘Show’ vs. ‘No-Show’ based on appointment status, and Cancelled appointments were removed. For each patient, the total number of appointments, number of no-shows, and no-show ratios, defined as number of no-shows divided by total number of appointments, were tracked during the follow-up period. The number of all cause inpatient admissions, emergency room visits, and total healthcare costs were measured during the 2-year follow-up period. Chi-squared and t-tests (\(\alpha=0.05\)) were used to assess whether the ‘Show’ and ‘No-Show’ groups differed significantly with respect to HRU and costs. **Results.** Table 1 shows the summary statistics. 1,770 (46.4%) patients had at least one ‘No-Show’ appointment (no-show group) in the follow up period, and 2,041 (53.6%) did not have a no-show (show group, i.e. all appointments were ‘Show’). The no-show group was slightly younger (42.2 vs. 43.5 years) compared to the show group (\(p=0.01\)). About one third of no-show patients were male. Compared to those who did not have any no-shows, no-show patients were 2.2 times more likely to utilize emergency room, and 1.6 times more likely to be hospitalized during the 2-year post-index period. Moreover, the two-year total cost of care was, on average, $10,057 higher among no-show patients compared to their show counterparts. Within the no-show group, we also explored the association between no-show ratio and ER and inpatient utilization. Patients with at least one inpatient admission had 42% rate of no-show to their appointments, compared to those without any inpatient admission who no-showed at rate of 37% (\(p=0.04\)). Similarly, no-show ratio of ER utilizers was 41% and higher than 38% for ER non-utilizers (\(p=0.15\)). **Conclusion.** In this study, we illustrated that missing a scheduled appointment is associated with increased HRU and higher costs in a cohort of BPD patients. The findings highlight the importance of ambulatory care visits, and policies designed to improve appointment adherence, especially in mental health populations. Future work will be multivariate analyses of association between appointment no-show and HRU and cost, adjusting for patient demographics, comorbidities, and social determinants of health.

**Table 1.** Summary Statistics. OR=Odds Ratio. CI=Confidence Interval

<table>
<thead>
<tr>
<th>Variable</th>
<th>Show N=2,041</th>
<th>No-show N=1,770</th>
<th>OR</th>
<th>OR 95% CI</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age at index (Mean, SD)</td>
<td>43.5</td>
<td>15.8</td>
<td>42.2</td>
<td>14.6</td>
<td>0.01</td>
</tr>
<tr>
<td>Male (N, %)</td>
<td>667</td>
<td>517</td>
<td>0.9</td>
<td>0.7, 1.0</td>
<td>0.02</td>
</tr>
<tr>
<td>Emergency Room - Yes (N, %)</td>
<td>1,522</td>
<td>1,535</td>
<td>0.9</td>
<td>0.7, 1.0</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td># of ER visits (Mean, SD)</td>
<td>3.6</td>
<td>6.4</td>
<td>2.2</td>
<td>(1.9, 2.6)</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Hospitalization - Yes (N, %)</td>
<td>810</td>
<td>908</td>
<td>1.6</td>
<td>(1.4, 1.8)</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Total cost (Mean, SD)</td>
<td>$32,238</td>
<td>$44,427</td>
<td>$42,295</td>
<td>$57,412</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

**References**

Development of a Clinical Question-Answering Corpus with Realistic Multi-Answer Challenges

Sungrim Moon, PhD\(^1\), Huan He, PhD\(^1\), Jungwei W. Fan, PhD\(^1,2\)
\(^1\)Department of Artificial Intelligence and Informatics
\(^2\)Kern Center for the Science of Health Care Delivery, Mayo Clinic, Rochester, Minnesota

Abstract: We generated a clinical question-answering (QA) corpus that involves realistic, multi-answer cases by converting the concept-relation annotations from the i2b2 Adverse Drug Event extraction challenge dataset. A total of 91,440 QA entries were derived, representing a diverse array of 1-to-0, 1-to-1, N-to-1, and 1-to-N relations between the question and answer(s).

Introduction: Clinician’s thought processes and decision-making often involve a cascade of questions and answers. Achieving human-like question-answering (QA) capability is highly regarded in artificial intelligence (AI). Medical QA research has therefore gained vital momentum over the past decade. A new generation of AI scientists is updating the “state-of-the-art” at a daunting pace almost every few months, if not weeks. One of the sought-after applications is to find the answer within a given document (a.k.a. reading comprehension), which enables patient-specific QA based on the information mentioned in the clinical text. As an essential component in AI engineering, QA training data determines the likelihood of success not only in terms of annotation quality but the fidelity of representing the target scenario. Along with other issues observed in existing medical QA corpora, the mainstream annotation approach knowingly simplifies the task into a one-answer-one-document scheme. The simplification has practically made development and evaluation easier for promoting the initial growth of the field. However, it is unrealistic in the sense that medical QA can naturally have multiple qualified answers within one document – sometimes all need to be captured to sufficiently answer a question. To address this gap, we created a clinical QA corpus that involves realistic multi-answer cases by converting the concept-relation annotations from an existing AI challenge dataset.

Methods: The 2018 i2b2 Adverse Drug Event (ADE) challenge\(^2\) was originally organized for extracting ADEs and various drug-related entities in clinical text. The dataset contains a gold standard annotation of 83,869 concepts and 59,810 relations in 505 discharge summaries. Given our primary interest in why-QA, we focus on generating QA pairs from the subset of annotated Reason-Drug relations. Each Reason-Drug relation has two arguments: one Reason concept and one Drug concept. Accordingly, we can generate a question from a Drug, e.g., “Why morphine was prescribed to the patient?” and its Reason “pain” can then be made the answer. Since there exist many-to-many relations across the Reason and Drug concepts in the dataset, we can derive the corresponding N-to-1 or 1-to-N QA entries as a logical result. In addition, by leveraging the fact that some Drug concepts do not have any Reason in the text, these no-reason instances allow us to generate unanswerable questions – a critical subset in QA challenges for evaluating AI system specificity. As another value-adding feature, we inject paraphrastic questions for enhancing the generalizability of the potentially trained systems.\(^3\)

Results & Discussion: The table below summarizes the QA categories and examples in our generated corpus. We obtained a total of 91,440 QA entries, of which half (51%) are unanswerable. The answerable ones include a substantial percentage of 1-to-1 (24%), N-to-1 (15%) and 1-to-N (10%) QAs, and each has more than 9,000 entries. The results are formatted into the widely adopted Stanford Question Answering Dataset (SQuAD) 2.0 JSON, except that we now allow each 1-to-N QA to have a list of reasons under the answer block. As the major contribution of this work, these diverse and sizable annotations represent more realistic train/test data for developing robust clinical QA systems that will transcend the conventional, limited single-answer task definition. We are in the process of engaging our domain experts to quality-audit the corpus and will share it with the research community.

<table>
<thead>
<tr>
<th>Category</th>
<th>Frequency (%)</th>
<th>Example</th>
<th>Processing method</th>
</tr>
</thead>
<tbody>
<tr>
<td>No reason annotated (1-to-0 QA)</td>
<td>46,278 (51%)</td>
<td>Mirtazapine 15 mg PO QHS [only the drug is mentioned but no reason is documented]</td>
<td>Make it an unanswerable QA entry</td>
</tr>
<tr>
<td>1 drug 1 reason (1-to-1 QA)</td>
<td>21,906 (24%)</td>
<td>The patient received morphine for pain as needed</td>
<td>Make into a 1-to-1 QA entry</td>
</tr>
<tr>
<td>N drugs 1 reason (N-to-1 QA)</td>
<td>14,004 (15%)</td>
<td>Hypertension: Severely elevated blood pressure. Started amlodipine, metoprolol, and isorbidie</td>
<td>Break into N separate 1-to-1 QA entries</td>
</tr>
<tr>
<td>1 drug N reasons (1-to-N QA)</td>
<td>9,252 (10%)</td>
<td>albuterol sulfate 90 mcg… Puff Inhalation Q4H for sob or wheeze</td>
<td>List the N reasons under answer block to form a 1-to-N QA entry</td>
</tr>
</tbody>
</table>

References

Characterizing Outpatient Portal Use among Women Receiving Obstetric and High-risk Pregnancy Care

Evan Morgan, BA, Shonda Vink, MPH, Pallavi Jonnalagadda DrPH MBBS, Priti Singh PhD, MS, Naleef Fareed, PhD, MBA
The Ohio State University, Columbus, OH

Introduction
Outpatient portal technology (OPP) allow patients to engage with their healthcare. For pregnant women, the level of engagement could have important implications for maternal and infant outcomes.1,2 There is a dearth in studies that characterize OPP use among pregnant women. Our academic medical center (AMC) implemented the OPP in 2011 system wide. The most frequently used OPP functions at our AMC are those that allow patients to access their personal health information (PHI), view and schedule appointments, and message their providers.3 Our study is one to the first to our knowledge that characterizes OPP use and identifies user profiles in order to explore how a historically understudied patient population engages with their healthcare electronically.

Methods
We built upon existing research using OPP server-side log files by executing a hierarchical clustering algorithm to group pregnant women based on their proportion of use for each OPP function.4 Of the 14,658 women with optional use of the OPP in our initial dataset, 7,664 women were considered active users of the OPP during their pregnancy if they used the OPP at least once within each trimester of pregnancy. Women who visited a Maternal Fetal Medicine (MFM) provider for a pregnancy were considered high-risk, while those who only visited an Obstetrics and Gynecology (OB/GYN) provider were considered having a normal pregnancy. Post-hoc analyses were performed using one-way ANOVA to further assess OPP use on key encounter and OPP use characteristics. For the cluster analysis, use of the following OPP functions was examined: Messaging (send/receive messages), Visits (manage appointments), My Record (access PHI), Billing (view bills, insurance information). We calculated proportions of use for each OPP function across a pregnancy and used these proportions as inputs for our cluster analysis.

Results
Four clusters were identified among all pregnancy episodes based on our clustering stopping rules and upon inspection of the cluster groups. The “Average Users” (AUs) cluster consists of those who primarily used the My Record and Visits functions. The “Prepared Engagers” (PEs) and “Schedulers” (SCs) focused their use on the Visits function, with use more common among SCs. Lastly, the “Intense Digital Engagers” (IDEs) used the Messaging and My Record functions most. The SCs cluster was not seen among normal risk pregnancies, whereas the same clusters were seen among those with a high-risk pregnancy. Post-hoc analyses revealed the IDEs cluster had the least sessions per pregnancy, while the clusters centered on use of the Visits function (i.e., SCs and PEs) had many sessions per pregnancy. This suggests that the Visits oriented clusters may be less focused in their use of the OPP functions in comparison to the other clusters. Likewise, SCs and PEs had many sessions per pregnancy. These women are also eager to engage with the OPP. Notably, clusters in the high-risk pregnancy group had more sessions per pregnancy, perhaps indicating the importance of maintaining a higher level of engagement with providers.

Conclusion
Characterizing OPP use among pregnant women demonstrates how specific patient populations might engage with their health. We were able to identify distinct cluster groups of OPP users among pregnancy groups. Future studies could explore whether such patterns of OPP use influence patient quality and clinical outcomes. For instance, OB/GYN or MFM teams may wish to know if they are retaining women who primarily use the Visits function, or if medication adherence is greater among those who primarily use the MyRecord function. Finally, this work can support studies of motivational themes of OPP use, which might help expand our understanding of OPFs as a beneficial tool in the effort to improve patient engagement.

References
Assessing CONCERN: Analysis of Application Log Files to Investigate the Utilization of a Clinical Decision Support Tool for Identifying Risky Patients

Amanda J. Moy, MPH, MA1, Kenrick D. Cato, RN, PhD,2,3 Christopher Knaplund, MPhil,1 Patricia Dykes, RN, PhD,2,4 Min Jeoung Kang,4 RN, PhD, Graham Lowenthal,1 Sarah Collins Rossetti, RN, PhD1,2
1Columbia University Department of Biomedical Informatics, NY, NY; 2Columbia University School of Nursing, NY, NY; 3Columbia University Irving Medical Center Department of Emergency Medicine, NY, NY; 4Brigham and Women’s Hospital, Boston, MA; 5Harvard Medical School, Boston, MA

Introduction
Communicating Narrative Concerns Entered by Registered Nurses (CONCERN) is a clinical decision support (CDS) tool that leverages data from the electronic health record (EHR) to make predictions about a patient’s risk for deterioration. The CONCERN risk score has demonstrated significantly better performance in predicting patient status when compared to other early warning systems.1 In this retrospective analysis, we were interested in examining whether application (app) log files were useful in uncovering usage patterns (e.g., link between risk and features accessed and risk and intensity of use, usage over time) among RNs and physicians (MDs) following its real-world implementation.

Methods
We extracted CONCERN’s app usage log file data from 6/2020 (i.e., go-live) to 1/2021. At the time, CONCERN was implemented in the acute care and intensive care units of a large Northeastern U.S. academic medical center; additional study details are described elsewhere.1 Log files comprised of all events (i.e., activity types) that RNs and MDs engaged in within CONCERN after launching it from the EHR patient list. We grouped log events into higher-level categories based on app features: clicking links to resources (i.e., contact us, FAQs, review the research, view the model, watch the video), predictive model driving factors (i.e., flowsheet comments, medication administration record, note content, note metadata and vitals) and viewing trend time window change (i.e., visualizations of longitudinal data). Log files displayed risk score (1=low to 10=high), score color (green=low, yellow=moderate, red=high), and time associated with each hourly update for each patient risk score. We restricted usage and trend analyses to 6,456 events for which time and risk data were complete. Session is defined as all events RNs and MDs engaged in per hourly score update.

Results
Among the 375 RNs and MDs included, the app was launched 1,325 times (38.9% were green, 47.5% yellow, 13.6% red) across 870 patients; 54 launches occurred within the same hourly risk score update. Across all events stratified by score color (Figure 1), predictive model driving factors were accessed at higher proportions (12.3-23.6%) compared to other categories, such as clicking links to resources (0.1-3.6%) or viewing trend time window change (0.7-5.1%). The association between event types (Figure 1) and score color (p<.001), and event counts per score update and score color (p<.001) were statistically significant; 81.5% of all RN and MD app sessions involved 0 to 5 events (Figure 2).

Discussion and Conclusion
The provision of CDS tools in clinical settings does not ensure uptake, and clinician resistance—such as lack of trust is a barrier.2 Evaluating usage patterns through app log files may be one approach to examining uptake of CDS tools among clinicians and may shed light on how to improve design and sustain usage. We harnessed information generated in CONCERN’s app log files to study utilization behaviors within the app and found that RNs and MDs were more likely to use features related to immediate patient care (i.e., predictive model driving factors) versus ancillary resources (i.e., clicking links to resources) and longitudinal data displays (i.e., viewing trend time window change) based on the app events captured; event types and score color and event counts per session and score color were statistically significant, but not attributable to increasing severity. Over 81% of the sessions represented involved 0-5 events. Future analyses will explore event duration, launch rate and how visual cues on model feature contribution impact usage.1

Acknowledgements: This work was supported by the U.S. National Library of Medicine (T15LM00707) and NINR (1R01NR016941-0).

References
An Example of Leveraging State Claims Data to Evaluate COVID-19 Policy Changes regarding Substance Abuse Disorder/Opioid Use Disorder Programs

Charles Mueller, PhD1, Johnna Beane1 Tsan-Yao Huang, PhD1, Thomas Bias, PhD1, Jennifer Mallow, RN, FNP-BC, PhD2

1West Virginia University Office of Health Affairs & School of Nursing, Morgantown, WV, USA

Introduction
As the state agency responsible for the administration of Medicaid benefits, the West Virginia (WV) Bureau of Medical Services (BMS) provides health insurance coverage to nearly one third of the population (N=564,000)1. In response to COVID-19, the WV BMS introduced new and authorized changes to its Medicaid policies to minimize human to human contact and ensure continuity of healthcare services2. It is anticipated that these policy changes will have an impact on service utilization, Medicaid spending, and health outcomes during and after the COVID-19 pandemic. In order to determine the effectiveness of these policies, a set of prioritized policies are being evaluated to provide timely information and to inform decision-making on maintaining policy changes post-pandemic.

Methods
A comprehensive list of WV Medicaid policy changes that were implemented in response to the COVID-19 pandemic were compiled along with documenting the contextual reasons for implementing the change. Next, the development of evaluation strategies for each policy change were identified as well as the proposed source of data. Medicaid claims data were examined to identify patterns and trends to provide insight about cost, service utilization and health outcomes.

The nature of conducting an evaluation during a continuing public health emergency requires a flexible approach to answer evaluation questions due to time and response constraints from affected populations. In addition, many of these evaluation questions were designed to be answered using sequential mixed methods designs, meaning that claims information will further be integrated with the results of a subsequent qualitative analysis regarding the same question.

Results
The evaluation plan consisted of 32 questions with a Medicaid claims analysis component. Due to space and time limitations, we present one example from our findings. In order to look at the impacts of BMS authorizing the use of telehealth flexibilities to replace face-to-face counseling/therapy requirements, we created a set of codebooks that allowed us to extract the following indicators: Number of counseling/therapy visits per MAT recipient; MAT recipients as a percentage of total Medicaid population; Proportion of MAT counseling/therapy visits occurring face-to-face vs. via telehealth; and MAT services as a percentage of total services billed by types of facilities. Presented above (Figure 1) is the result we obtained for indicator related to telehealth utilization showing the monthly relative percent of MAT counseling/therapy services rendered via telehealth from January 2019 to October 2020.

Discussion
While the results are preliminary, Figure 1 shows the percentage of counseling/therapy services occurring via telehealth drastically increased after the introduction of this policy with peak utilization occurring in April 2020 at around 43%. This began to drop in May 2020 and has since ranged between 20-30% per month, suggesting that a substantial portion of the MAT population has been able to continue their counseling/therapy service requirements as a result of telehealth options. Starting on April 30th, 2020 the Governor of WV began removing the strictest COVID-19 precautions increasing the ability for face-to-face interactions3. Taken together the data suggests that even with increasing ability to meet face-to-face, some desire to use telehealth remains.

Conclusion
The ability to rapidly translate scientific data into easily comprehensible forms during a public health emergency helps create a decision feedback loop based on the best available information for policymakers. We demonstrate that informatic processes allow for systematic evaluation of health data (e.g. Medicaid claims data) in a way that is useful for policymakers. In the case of this project, we were able to provide WV BMS with direct feedback regarding how policy changes to SUD/OUD programs are impacting the target population during a public health emergency. This data demonstrates the successful use of telehealth modalities on delivering MAT counseling/therapy visits to address a critical need for a population of most importance in the state of WV. This information has policy implications as they consider ways to improve healthcare service delivery in rural areas of WV where in-person counseling services may not always be an option.

References
Title: Incorporating Health Equity into the Evaluation of Health Informatics Electronic Health Record Tickets: Design and Initial Findings

Authors: Anoop Muniyappa, MD, MS*; Colin Purmal, MD*; Jennifer Corbell, BSN, RNii; Susan Chim, DHA, MATiii; Kendall Gross, PharmDiv; Amy Kangwankij, BSN, RNii; Katie O’Connor, BSN, RN, OCNii; Kay Burke, MBA, BSN, RN, NE-BCii; Andrew Auerbach, MD, MPHii

*co-first author; 1Department of Medicine, University of California, San Francisco, CA; 2Department of Nursing Informatics, University of California, San Francisco, CA; 3Department of Health Informatics, University of California, San Francisco, CA; 4Department of Pharmacy, University of California, San Francisco, CA

Background: Health informatics interventions have the potential to both improve and worsen health equity.1,2 While frameworks have been proposed to guide assessment of the impact of health technologies on equity,3,4,5 few have demonstrated a practical and reproducible method to evaluate electronic health record (EHR) tools with regard to health equity.

Objectives: To describe feasibility and early experience with a system to evaluate ordersets, documentation tools, and decision support tools with regard to health equity.

Methods: The UCSF Health APeX Clinical Content Committee, the governance group responsible for overseeing provider-facing EHR tools, created an equity scoring rubric that included: 1) Classification of tickets as beneficial effect on health equity, neutral, or needing improvement, 2) Assessment of impact on potentially affected patient groups using a Likert scale (1: very low, 2: low, 3: intermediate, 4: high, 5: very high), and 3) Description of the benefit or suggestions for improvement. Using these criteria, we piloted our schema by reviewing 20 EHR change tickets, calculated descriptive statistics, and performed a thematic analysis.

Results: Three informaticists reviewed 20 EHR change tickets using our internal rubric. Of these, 6 were scored as beneficial, 13 neutral, and 1 needing improvement. Beneficial ticket impact scores ranged from 1 to 4 with an average score of 2.83, suggesting low to intermediate impact. The ticket flagged for improvement had an impact score of 3, suggesting intermediate impact. Descriptions of potentially beneficial tickets focused on improving care for conditions inequitably distributed across the population by race, ethnicity, and/or socioeconomic status, standardizing care for various conditions, and expediting mental health care. The ticket flagged for improvement provided helpful patient instructions but was only developed in English.

Conclusions: Establishing criteria to evaluate the impact of EHR change tickets on health equity represents a novel way to identify potential disparities in EHR enhancements, and can promote health equity through improved EHR governance. Our next steps include incorporating health equity scoring into enhancement prioritization, growing our list of examples representing how health equity is impacted by EHR changes, and evaluating interobserver variability in scoring.

Addressing Challenges and Strategies for Virtual Recruitment for Longitudinal Studies

Annie Myers, MA1, Meghan Reading Turchioe, PhD, RN1, Sabrina Mangal, PhD, RN1, Leslie Park, BS1, Lisa Grossman Liu2, Ruth Masterson Creber, PhD, RN1

1Department of Population Health Sciences, Division of Health Informatics, Weill Cornell Medicine, New York, NY; 2Department of Biomedical Informatics, Columbia University, New York, NY

Introduction
Since March 2020, the COVID-19 pandemic has presented unique challenges and interrupted traditional in-person recruitment strategies.1 For the past year many researchers have turned to remote virtual recruitment. Remote recruitment requires that participants have access to technology, while in-person recruitment does not. This poses a significant challenge for vulnerable populations who may not have access to email or a smartphone. We discuss a few recruitment tools and strategies that have helped us continue our patient recruitment when in-person recruitment was no longer possible.

Methods
We started virtual recruitment for vulnerable patients with heart failure in March 2020. For approaching eligible patients, we utilized flyers and phone scripts, and once enrolled, data collection was supported by the Research Electronic Data Capture (REDCap) platform.2 Flyers were designed by author LGL and were placed in the participating clinician’s offices (Figure 1). We found that it was more effective for clinician’s to directly give these flyers to patients compared to indirectly having them in the waiting room. After being given the flyer, patients were called by a research coordinator using specific phone scripts. The phone scripts included the name of their provider, explained the purpose of the study, and that it was all completed online through REDCap. REDCap is a secure, web-based software platform designed to support data collection and data tracking. We utilized REDCap’s ‘Alerts & Notifications’ feature which allows researchers to construct alerts and send customized notifications to patients for follow-up. These alerts can be scheduled to automatically send at specific times for longitudinal projects (i.e., after 2 weeks, 4 weeks, etc).

Results
Although some patients had difficulty with Internet connection, the majority were able to easily complete the enrollment process and online surveys online. Patients who reported monitoring their symptoms or using technology often had higher rates of completing follow-up surveys through email or text.

Conclusions
There are many challenges to virtual and remote recruitment. Virtual recruitment is undoubtedly more difficult, especially for vulnerable older populations. However, the tools used for this project’s virtual recruitment have allowed us to continue recruiting patients at a slower but steady rate during the COVID-19 pandemic. This recruitment is supported by NINR of the NIH under Award Number R00NR016275 (PI: Masterson Creber).

References
Developing Framework for Healthy Coping Feedback Messages Evaluation from Diabetes Mobile Apps

Ploypun Narindrarangkura, MS, MD1, Qing Ye, MS, PhD1, Suzanne A. Boren, MHA, PhD1,2, Uzma Khan, MD3, Eduardo J. Simoes, MD, MSc, MPH1,2, Min Soon Kim, PhD1,2

1Institute for Data Science and Informatics; 2Department of Health Management and Informatics; 3Department of Medicine, University of Missouri, Columbia, MO, USA

Introduction

The American Diabetes Association (ADA) has recommended that all individuals with diabetes should receive diabetes self-management education and support (DSMES) throughout the treatment process. AADE7 Self-Care Behaviors® (AADE7®) is a validated and widely accepted DSMES principle. AADE7® has seven principles: Healthy Eating, Being Active, Monitoring, Taking Medication, Problem Solving, Reducing Risks, and Healthy Coping. In this study, we focused on Healthy Coping, a key principle that enables people with diabetes to achieve self-care goals. Healthy Coping supports monitoring diabetes-related measures, physical exercise/activity, and mood of people with diabetes. Diabetes mobile applications with feedback messages can be used to promote Healthy Coping. Feedback messages should be designed based on behavioral change theories because a feedback message is to allow people to reflect on their progress and remind them of their goals. We aimed to analyze Healthy Coping-related feedback messages from diabetes mobile apps against the theoretical framework based on validated behavioral change theories.

Methods

There was a lack of a framework to evaluate feedback messages from diabetes mobile apps. We reviewed the literature to identify the most suitable model that we could use for evaluating feedback messages. A framework was developed based on several validated behavioral change theory-based models (Figure 1). The framework is composed of three domains and three dimensions. The three dimensions are timing, intention, and content of feedback messages. The content dimension has two subdimensions of feedback purpose and feedback response. For the feedback purpose, there are seven types of purpose: warning, suggestion, self-monitoring, acknowledging, reinforcement, goal setting, and behavior contract. The feedback response is divided into real-time and delayed responses. Real-time messages provide instant information when user input data, while delayed messages provide information at a later time after user input data. Then, we searched diabetes apps from iTunes and Google Play stores, and entered a range of values on three Healthy Coping domains: 1) diabetes-related measures including blood glucose, blood pressure, HbA1c, weight, 2) physical exercise/activity, and 3) mood to generate feedback messages. We analyzed the feedback messages against three dimensions of timing, intention, and content.

Results

We selected eligible 156 out of 1,749 diabetes mobile apps that generated a total of 473 feedback messages. The majority of messages were related to blood sugar (219, 46%) and followed by mood domain (128, 27%). For the timing dimension, most of the messages were user-initiated (336, 71%). Most of the messages across all domains had neutral intentions (358, 76%), simply confirming user activity and asking to continue the activity. More than 75% of messages among all Healthy Coping domains were real-time responses. Only feedback messages on blood sugar under diabetes-related measures and mood domains encompassed all seven feedback purposes under the content dimension.

Conclusion

We found that there was an unbalanced distribution of feedback messages across healthy coping domains and feedback dimensions. It is important that feedback messages be structured around the dimensions of the behavioral theory-based framework to promote behavior change. Our framework can be extended to evaluate feedback messages for other AADE7® principles and other chronic diseases because the backbone of the model used in this study is evidence-based.
Machine Learning Implementation Results in Reduced Variable Cost of Acute Care Visits During Radiotherapy: SHIELD-RT Cost Analysis

Divya Natesan MD1, Samantha M. Thomas MS2, Eric Eisenstein DBA3, Neville C. W. Eclov PhD1, Nicole H. Dalal MD4, Sarah J. Stephens MD4, Mary Malicki MSN ACNP1, Stacey Shields ANP-BC1, Alyssa Cobb, RN BSN1, Yvonne M. Mowery MD PhD1, Donna Niedzwiecki PhD2, Jessica D Tenenbaum PhD3, Manisha Palta MD3, Julian Hong MD MS4
1Duke University Medical Center, Department of Biostatistics and Bioinformatics, Durham, NC; Duke University Medical Center, Department of Radiation Oncology, Durham, NC; Duke University, Durham, NC3; University of California, San Francisco, CA4

Introduction: SHIELD-RT was one of the first prospective randomized studies (PMID: 32886536) to utilize ML in an interventional clinical application. SHIELD-RT implemented a ML algorithm to identify patients who were at high risk of acute care visits (ER visits or hospitalizations) during radiotherapy (RT). High risk patients who were randomized to receive supplemental visits during treatment had reduced rates of acute care utilization (12%) compared to standard care (22%). We hypothesized that acute care variable healthcare costs (impacted by patient volume and vary with usage) would also be reduced in those who received supplemental visits.

Methods: Patients who initiated RT from 1/7/2019-6/30/2019 at a Duke University Hospital were evaluated during the first week of treatment by a ML algorithm to identify high risk patients (>10% risk of acute care visit during RT). High risk patients were prospectively randomized to receive either standard weekly (S) or intervention of twice weekly (TW) visits during RT. Variable cost data (excluding fixed costs) associated with acute visits were obtained and compared between patients who underwent S or TW evaluations. Missing cost data were imputed using disease related groups. Variable costs are presented as mean (standard deviation) and were compared between arms with non-parametric Wilcoxon Rank Sum tests. Intervention costs were not included in this analysis.

Results: 311 high risk courses were identified and randomized to receive either S (n=157) or TW (n=154) evaluations during RT. 85 patients (S: 51; TW: 34) had 121 distinct acute care visits (S: 74; TW: 47). Patients in the TW evaluation arm had fewer hospitalizations (29 vs 41) and ER visits (18 vs 33) than those in the S arm. Actual cost data was available for 102 visits and imputed for 19 outside hospital visits. Variable costs associated with acute visits was lower in the TW arm ($156,066) compared to the S arm ($349,260). Mean cost per patient associated with acute visits was also lower in the TW arm ($1013, SD $3201) compared with the S arm ($2225, SD $6243) (p=0.02).

Table 1. Mean Variable Costs of Acute Care Per Patient, By Revenue Center

<table>
<thead>
<tr>
<th>Revenue Center</th>
<th>Standard (N=157)</th>
<th>Twice Weekly (N=154)</th>
<th>Difference in Means (95% CI)</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inpatient ($)</td>
<td>2037 (6044)</td>
<td>927 (3054)</td>
<td>1110 (39, 2182)</td>
<td>0.03</td>
</tr>
<tr>
<td>Room ($)</td>
<td>1177 (3688)</td>
<td>481 (1662)</td>
<td>696 (56, 1337)</td>
<td>0.04</td>
</tr>
<tr>
<td>Labs/Radiology/Diagnostics ($)</td>
<td>257 (670)</td>
<td>147 (406)</td>
<td>110 (-14, 234)</td>
<td>0.05</td>
</tr>
<tr>
<td>Pharmacy ($)</td>
<td>369 (1379)</td>
<td>198 (871)</td>
<td>171 (-87, 428)</td>
<td>0.04</td>
</tr>
<tr>
<td>Surgery/Anesthesia/Cardiology/Respiratory/Rehabilitation/Speech ($)</td>
<td>234 (1125)</td>
<td>101 (469)</td>
<td>133 (-60, 326)</td>
<td>0.04</td>
</tr>
<tr>
<td>Emergency Services ($)</td>
<td>187 (742)</td>
<td>86 (231)</td>
<td>101 (-22, 224)</td>
<td>0.10</td>
</tr>
</tbody>
</table>

Discussion: Acute care visits during RT are a frequent and costly occurrence for high risk patients. ML-directed evaluations decreased variable costs of acute care. Variable costs were reduced across revenue centers, with the largest difference related to inpatient room costs. These are the first data to report on the variable cost implications of ML implementation in a clinical setting. Analyses incorporating physician and intervention costs are underway.
Diachronic Analysis of the Evolution of COVID-19 Scientific Literature

Denis Newman-Griffis, PhD1, Venkatesh Sivaraman2, Adam Perer, PhD2, Eric Fosler-Lussier, PhD3, Harry Hochheiser, PhD1

1University of Pittsburgh, Pittsburgh, PA; 2Carnegie Mellon University, Pittsburgh, PA; 3The Ohio State University, Columbus, OH

Introduction The COVID-19 pandemic has led to an unprecedented global research effort. New findings emerge constantly and are swiftly disseminated, and scientific understanding of the pandemic has evolved at a rapid pace. The COVID-19 Open Research Dataset (CORD-19),1 an actively-maintained collection of scientific literature related to COVID-19, indexes tens of thousands of new publications and preprints each month, far more than any human reader can consume. Computational analysis of changes in the literature is thus required to understand trends in how scientific understanding has developed during the pandemic, and identify patterns suggesting promising new research directions.

We developed TextEssence, a text analysis workflow with an interactive, web-based interface to explore diachronic shift in language use, as well as other semantic differences between text collections. TextEssence reflects key shifts in the COVID-19 literature between March-October 2020, and lays the foundation for further knowledge discovery in the evolution of COVID-19 research.

Methods Our analysis leverages concept embeddings to capture patterns in how medical concepts are discussed in CORD-19: i.e., which concepts play similar roles in scientific writing (such as treatments for COVID-19 symptoms). Our workflow, illustrated in Figure 1, had three phases. We used the JET software2 to train embeddings of SNOMED CT concepts from snapshots of CORD-19 containing all new publications indexed in each month from March to October 2020. These embeddings were filtered to the concepts with stable representations across multiple replicates. We built a web-based interface for qualitative analysis of the resulting embedding spaces, providing three functions: (1) Browsing a set of embeddings; (2) Inspecting how a particular embedding changes over time; and (3) Comparing how a pair of concept embeddings changes relative to one another over time.

Results The embeddings of key concepts related to COVID-19 changed across the months of CORD-19 data in ways that reflected the broader evolution of scientific understanding around the pandemic. Anosmia moved from being embedded with other otolaryngological concepts in early months to associations with other symptoms such as nausea and vomiting in October 2020, reflecting its earlier emergence as a key diagnostic variable. The corticosteroid Dexamethasone, used throughout the pandemic to treat severe COVID-19 symptoms, picked up Ruxolitinib as a nearest neighbor in September 2020, reflecting a spike in the literature on similar use of ruxolitinib for severe symptom relief. Hydroxychloroquine added nearest neighbors such as hospital-acquired infections and respiratory failure over time, reflecting the body of studies finding negative or harmful results of its usage for COVID-19 treatment.

Conclusion These results indicate that concept embeddings provide a useful lens for studying shifts in the COVID-19 scientific literature. Our work provides a strong foundation and reusable implementation for exploring the use of medical concept embeddings for discovery of new knowledge and relationships over time. The results of our CORD-19 analysis are available at https://doi.org/10.5281/zenodo.4432958, and the TextEssence system is available from https://github.com/drgrippis/text-essence. A screencast of the system is available at https://youtu.be/1xEEfsMwL0k.

References

Qualitative Coding Framework for Analyzing Alert Notes from the Telehealth Intervention Program for Seniors (TIPS)

Phuong Nguyen, MPH.¹, Hyung Wook Choi, M.S.², Melody K Schiaffino, MPH, Ph.D.¹, Zhan Zhang, Ph.D.³, Jina Huh-Yoo, MHCI, Ph.D.²

¹San Diego State University, San Diego, PA, USA, ²Drexel University, Philadelphia, PA, ³Pace University, New York, NY, USA

Introduction

Telehealth brings opportunities to reduce hospital readmission and emergency visits and improve chronic illness care for older adults. Alerts from abnormal monitoring during telehealth participation can generate alert fatigue for clinicians.¹ Although alert fatigue has been extensively discussed in the context of electronic health records, little is known about telehealth and the efficacy of alerts. Alert notes, which describe nurse’s detailed notes about an alert, contain a larger context than medical record data to delineate whether an alert was false, how the problem was resolved, and any recorded steps moving forward for the patient. However, much of the literature depends on the quantitative analysis of patient’s medical records. In this poster, we introduce a qualitative coding framework developed for the alert notes generated from the Telehealth Intervention Program for Seniors (TIPS)—a tri-state community-based telehealth program implemented in the states of New Jersey, Pennsylvania, and New York (N=17 sites, N=1,878 participants, N=24,933 alert notes, 2011-2019). The framework enables mapping the entire workflow and sequence of events leading from alert generation to nurse’s phone call attempts to patients and next steps. We developed keyword-matching information extraction techniques that represent the workflow of alerts and nurse calls with patients based on 352 manually annotated alert notes. We discuss the scalability of this method in the large-scale analysis of alert notes and use of the framework for future analyses.

Methods

A research team with gerontology, public health, telehealth, and text processing expertise generated the workflow model shown in Figure 1 to map the information documented in the alert notes regarding the nurse’s response to each alert. The workflow model presents multiple sequences of events that begin with an alert to the nurse’s call to the patient and resulting plan for the patient. We developed a coding framework based on the workflow model, and keyword matching information extraction approach building on manually annotated alert notes for scalable analysis of the alert notes. We iteratively refined the regular expression that would capture the notes that represent each code. For instance, we used keywords “equipment”, “issue_machine”, “malfunction”, “not_work_properly” to capture notes that present the code 1a: ‘technical error’. We then tested the keyword match results by randomly selecting 400 mutually inclusive alert notes (20 identified as false, 20 identified as negative) for 20 codes, resulting in analyzing 762 alert notes in total for precision and recall analysis.

Results and Conclusion

Using the keyword matching approach, we identified a total of 23,856 alert notes identified as belonging to at least one of the 20 codes (Precision: 0.89, Recall: 0.91, N=762). The high-performance results show the scalability of this approach. The resulting coded model can be cross analyzed with patients’ medical records, which can generate numerous predictive models, correlations, and appropriate triaging of false alerts. Furthermore, the TIPS program has a high proportion of racial and ethnic minority participants and those with limited English proficiency—all with known barriers to healthcare utilization. These socio-economic factors can be mapped with the codes to aid identifying facilitators and barriers to making successful monitoring that lead to increased health equity and continued retention.

References

Identifying Posts from Reddit Talking About ADRD Caring Experiences Using Machine Learning

Congning Ni1, ME; Jonathan A Hughes2; Andrew K Nam1; Zhijun Yin1,3, PhD

1Vanderbilt University, Nashville, TN, USA,
2Lipscomb University, Nashville, TN, USA
3Vanderbilt University Medical Center, Nashville, TN, USA

Background Alzheimer’s disease and related dementia (ADRD) impair memory, thinking processes, and mainly occur in older adults1. It is estimated that the caregivers of people living with ADRD provide more than 18.5 billion hours of unpaid care for their family members in the United States every year2. Therefore, it is essential for the society to understand and respond the challenges faced by these caregivers3. With the rapid development of internet, many ADRD caregivers began to seek support or share caring experiences on various social media platforms or online health communities. However, given the massive, noisy online data, it is necessary to classify the posts with ADRD caring-related content before further analysis.

Objective We aimed to build efficient classification models to identify online posts about ADRD-caring experiences from Reddit, a social news aggregation, web content rating, and discussion website.

Methods Firstly, our data consists of all the submissions collected through the pushshift.io Reddit API by using both 1) ADRD-related keyword searching (“alzheimer”, “alzheimers”, “demen” and “dementias”) and 2) collecting all the submissions from ADRD related subreddits (r/Alzheimer, r/Alzheimers, r/dementia, and r/AlzheimersGroup) from 2012 to 2020. Then, we combined the two datasets and drop the duplicates. Next, we randomly sampled 5,000 submissions and recruited four annotators for annotating whether each of these submissions was related to ADRD care or not. Finally, based on the annotated results, we built multiple machine learning models for binary classification. We embedded the submissions with various feature extraction strategies, including using vectorization of Linguistic Inquiry and Word Count (LIWC), Term Frequency Inverse Document Frequency (TFIDF), Word Count, and N-grams (counting sequences of characters, n ∈ {2,5}). We applied three common classification models: a Logistic Regression (LR), a Random Forest Classifier (RFC) and a K-Nearest Neighbors (KNN) classifier. Additionally, we also tested three deep learning-based models, BERT fine-tuning, LSTM (one-hot representation) and LSTM (word2vec representation), to enhance model performance. For each model, we applied 10-fold cross-validation to select the optimal hyper-parameters, and evaluated model performances based on accuracy, precision, recall, f1 and AUC score.

Results 68,868 submissions were collected in the initial dataset. Within the 5,000 annotated submissions, 1,914 (38.28%) were labeled as positive, which were related to ADRD caring. Table 1 presents the AUC score of all the traditional models we trained. Table 2 shows the AUC of deep learning models. Note that the AUC of the BERT fine-tuning model was 0.938, which out-performed all other models.

Conclusion In this project, we built classification models to identify ADRD-caring-related posts from massive Reddit online posts. We demonstrated that these classifiers could efficiently identify ADRD-caring related information, which could be used for further analysis of topics shared by caregivers of ADRD patients on social media. Our future research will focus on 1) improving model performance and applying the model with the best performance for ADRD-related posts collection; 2) using the same strategy to other online communities of ADRD caregivers (e.g., AlzConnected), to obtain the related discussion in a broader range; 3) conducting content analysis to the data we collect to understand the needs of ADRD caregivers.

References

<table>
<thead>
<tr>
<th>LIWC</th>
<th>TFIDF</th>
<th>Word Count</th>
<th>N-grams</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR</td>
<td>0.801</td>
<td>0.818</td>
<td>0.780</td>
</tr>
<tr>
<td>RFC</td>
<td>0.838</td>
<td>0.841</td>
<td>0.833</td>
</tr>
<tr>
<td>KNN</td>
<td>0.700</td>
<td>0.699</td>
<td>0.594</td>
</tr>
</tbody>
</table>

Table 1. Traditional models’ average AUC score. The standard deviations are all under 0.025.

<table>
<thead>
<tr>
<th>LSTM One-Hot</th>
<th>LSTM Word2Vec</th>
<th>BERT Fine-Tuning</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.938</td>
<td>0.706</td>
<td>0.754</td>
</tr>
</tbody>
</table>

Table 2. Deep learning models’ average AUC score. The standard deviations are all under 0.019.
Evaluating the Scope of Collaboration Among Primary Care Teams through Electronic Asynchronous Communication

Arianna E. Nimocks, Bryan D. Steitz, PhD, Adam Wright, PhD
Dept. of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN

Introduction
Electronic health record (EHR)-based asynchronous communication (Inbasket messaging) has emerged as a primary means of communication and care coordination among multidisciplinary care teams. Within the ambulatory setting, primary care providers (PCPs) are often tasked to coordinate patient care across numerous specialties and care teams. Previous work has found that highly connected care teams had lower costs\(^1\) and better outcomes\(^2\). Analyzing the scope of communication and connectivity between care team members involved in asynchronous messaging affords key insights into opportunities to improve collaboration between clinical teams. In this study, we conducted a social network analysis to evaluate communication patterns among all members of the care team who communicated with PCPs, including physicians, clinicians, nurses, and administrative staff.

Methods
We extracted secure clinical message logs from the Epic EHR corresponding to each message viewed or sent by an adult PCP at the Vanderbilt University Medical Center between January 1, 2019 and June 1, 2019. We identified PCPs as any physician, physician assistant, or nurse practitioner who was scheduled for independent patient appointments in ambulatory general internal medicine or family medicine clinics. Message log data included a thread identifier, message timestamp, action (view or send), a unique employee identifier, and employee role for each messaging instance. From these logs, we calculated undirected edges by taking pairwise combinations of all individuals involved in a thread. For single messages that were not part of a larger thread, we indicated a tie between the sender and readers. We calculated descriptive messaging statistics and evaluated PCP relationships by employee role.

Results
During our study period, there were 1258 PCPs who communicated about 225161 patients through 1227813 messages – an average of 334 patients (SD=297) and 990 messages (SD=1016) per PCP. In Table 1, we show PCP relationship statistics. PCPs shared 57.4% of threads with only the top 5% of employees involved in their messages. The majority (64.2% and 62.1%) of PCP relationships shared fewer than three patients and three message threads, respectively.

<table>
<thead>
<tr>
<th></th>
<th>Administrative</th>
<th>Clinical Employee</th>
<th>PCP</th>
<th>Non-PCP Physician</th>
<th>Other</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Relationships (%)</td>
<td>1249 (15.1)</td>
<td>2969 (35.8)</td>
<td>1145 (13.8)</td>
<td>1093 (13.2)</td>
<td>1831 (22.1)</td>
<td>8287</td>
</tr>
<tr>
<td>Median Relationships per PCP</td>
<td>29</td>
<td>33</td>
<td>21</td>
<td>9</td>
<td>13</td>
<td>104</td>
</tr>
<tr>
<td>Number of Shared Patients (%)</td>
<td>92643 (41.1)</td>
<td>165093 (73.3)</td>
<td>49376 (21.9)</td>
<td>51573 (22.9)</td>
<td>102406 (45.5)</td>
<td>225161</td>
</tr>
<tr>
<td>Median Shared Patients per PCP</td>
<td>68</td>
<td>140</td>
<td>26</td>
<td>47.5</td>
<td>81</td>
<td>248</td>
</tr>
<tr>
<td>Number of Shared Threads (%)</td>
<td>162328 (20.0)</td>
<td>462661 (56.9)</td>
<td>81491 (10.0)</td>
<td>140996 (17.3)</td>
<td>217904 (26.8)</td>
<td>813577</td>
</tr>
<tr>
<td>Median Shared Threads per PCP</td>
<td>91</td>
<td>228</td>
<td>31</td>
<td>72</td>
<td>111</td>
<td>495</td>
</tr>
</tbody>
</table>

Conclusion
Improving care coordination among primary care team members has important cost and care quality implications. Our social network analysis found that most messages sent or received by PCPs involve a small “core” group of employees who commonly communicate, and that the majority of communicative relationships involved fewer than three message threads. These results suggest an opportunity to co-locate individuals who are distributed across the organization but work together closely to allow for more synchronous communication. Future work will examine how these collaboration patterns affect PCP workflow, with the goal of reducing unnecessary tasks.

References
Transforming Data Capture for Clinical Care and Research: ePRO Adverse Event Reporting and Quality of Life in a Phase 2 Breast Cancer Study

Anna Northrop¹, Michelle Melisko, MD¹, Garry Peterson¹, Laura Sit¹, Anika Christofferson¹, Ebunoluwa Olunuga¹, Ananya Mittal¹, Adi Goldman¹, Ronak Ahir¹, Mitra Rocca, PhD², Lisa Weiss, MBA³, Elizabeth Prager⁴, Cal Collins⁴, James Palazzolo, MS⁵, Laura Esserman, MD, MBA¹, Adam Asare, PhD¹,³, Amrita Basu, PhD¹

¹University of California, San Francisco, CA, ²U.S. Food & Drug Administration, Silver Spring, MD, ³Quantum Leap Healthcare Collaborative, San Francisco, CA, ⁴OpenClinica, Waltham, MA

Introduction: Here we describe the development of a system for monitoring patient-reported adverse events (AE) and quality of life (QOL) using electronic patient reported outcomes (ePRO) instruments in the Investigation of Serial studies to Predict Your Therapeutic Response with Imaging And moLeCular analysis (I-SPY 2 TRIAL), a phase II clinical trial evaluating neoadjuvant treatments for locally advanced breast cancer. I-SPY 2 randomizes patients to 12 weeks of a drug treatment regimen, followed by four biweekly cycles of Adriamycin Cyclophosphamide (AC) and surgery.

Methods: An ePRO survey series was designed to assess QOL and AEs in I-SPY 2 participants using a maximum of 126 validated, branching logic questions from the Patient Reported Outcomes Measurement Information System (PROMIS®) Health Measures and the National Cancer Institute’s Patient Reported Outcomes – Common Terminology Criteria for Adverse Events (PRO-CTCAE™) instruments.

We developed rules-based logic to build an automated scheduler and survey delivery platform using the OpenClinica (OC) electronic data capture system. We piloted the ePRO system at the University of California, San Francisco (UCSF) to evaluate feasibility and usability. For each patient who consented to screen for participation in I-SPY 2 at UCSF, the clinical research coordinator (CRC) linked the patient’s email and subject ID in the OC Participate™ module. This information was encrypted in the study database. When the patient consented to treatment, the CRC entered key clinical dates in the OC Scheduler case report form including: (1) date of screening consent, (2) start of drug regimen, (3) start of AC cycles, and (4) date of surgery. This prompted the system to automatically schedule and send ePRO surveys to the patient at a predetermined frequency based on the I-SPY 2 treatment schedule. Automated reminders were also put in place to encourage patient participation and increase compliance.

Results: The UCSF ePRO pilot began in September of 2020. Over nine months, 45 I-SPY 2 patients were accrued (average age of 43.8 years), whose interactions with the ePRO system informed design modifications and technologic iterations. Of the patients who received a baseline ePRO survey at UCSF, the completion rate was 75.9%.

CRCs monitored data completeness using the OC Participate™ module and a custom report that was built to display the status of each participant’s survey history in real time. These tools were designed to enable CRCs and clinicians to easily identify patients who had not completed their ePRO survey prior to their visit, so they may be provided with a tablet computer with internet connectivity to complete the survey in the clinic. After introducing tablets into the workflow at UCSF, patient completion of the baseline survey increased from 75.9% to 80%. To increase the accuracy of reporting, a CRC training guide was developed, which integrated detailed workflow diagrams and technical solutions for cases that deviated from the traditional I-SPY 2 patient journey.

Discussion: The implementation of ePRO within I-SPY 2 has the potential to improve the efficiency and accuracy patient-reported data collection for clinical trial research, while integrating learning as a byproduct of care. As a result of the refinements made during the development and pilot phase of the project, the ePRO system was implemented at 21 additional I-SPY 2 sites in June of 2021. We will seek to demonstrate the extensibility and scalability of the ePRO platform at these sites. Future iterations will integrate updated workflows as well as SMS capabilities and patient educational materials to increase survey completion rates.
Consumer Perception of Smart Speaker-Based mHealth Tool
Hyunkyoung Oh, PhD, RN1, Min Sook Park, PhD1, Youngjoo Cho, PhD2
1University of Wisconsin-Milwaukee, Milwaukee, WI; 2Konkuk University, Seoul, Korea

Introduction
Approximately 60% of Americans have at least one chronic condition.1 Self-management (SM) is a promising strategy for managing chronic health conditions in daily lives through identifying challenges and solving problems associated with their own illness.2 Although mHealth apps have been developed to support SM, most mHealth apps have not been used routinely by patients. Barriers to use include complicated and inconvenient ways to input and output data, as well as unreliable health information and data interpretation.3-4 The use of smart speakers has rapidly increased due to their ease-of-use and convenient features with advanced voice assistant and increased demand for smart devices.5 However, we do not clearly know patients’ acceptance and adoption of a novel smart speaker-based mHealth tool yet. Recognizing end-user needs and potential barriers to the use of technology is the most important key to develop a successful technology-based infrastructure to deliver a clinical intervention.6 The purpose of this study was to identify challenging SM behaviors and useful features of mHealth tool on a smart speaker platform.

Overview of methodology
A survey was applied using Qualtrics. We recruited individuals who are 18 years old and older living with at least one metabolic health condition (i.e., diabetes, hypertension, obesity) in the US through a Facebook advertisement. The survey asked what are most difficult SM behaviors. In the survey, we provided a brief description of our proposed Voice-activated Self-monitoring (VoIS) application on a smart speaker platform to know users’ perception and use intention to this tool, which has three types of features: self-monitoring, clinical decision support, and shared decision-making features (detailed features are reported in Table 1). The VoIS is able to connect to other smart devices (e.g., smartphone, tablet) thus, it can be used on multiple platforms which increases users’ ease-of-use and accessibility. The survey also asked what are most useful features after providing this description. Descriptive statistics were used for demographic characteristics. Mean rank values (MRV) were calculated to identify the order of priorities. Low MRV indicates a higher priority.

Results
A total of 509 survey responses were analyzed. The mean of age is 40.17 years and 66.2% are male. The majority have Bachelor’s degree (81.1%) and employed (71.9%). More than half are non-Hispanic White (63.5%) and 11% have more than two conditions. Over 70% reported that they would use the proposed mHealth tool continuously if they had it. Table 1 shows the most challenging SM behaviors and the useful feature of the proposed mHealth tool. Table 1. Rank order of challenging behaviors and useful features.

<table>
<thead>
<tr>
<th>Challenging SM behavior</th>
<th>MRV</th>
<th>Useful feature</th>
<th>MRV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monitoring and tracking health indicators</td>
<td>2.77</td>
<td>Monitoring and tracking health indicators</td>
<td>2.47</td>
</tr>
<tr>
<td>Taking medication</td>
<td>3.35</td>
<td>Interpreting recorded health indicators</td>
<td>3.63</td>
</tr>
<tr>
<td>Healthy diet</td>
<td>3.63</td>
<td>Smart speech reminder</td>
<td>4.2</td>
</tr>
<tr>
<td>Physical activity</td>
<td>4.42</td>
<td>Communication with health care providers</td>
<td>4.91</td>
</tr>
<tr>
<td>Stress management</td>
<td>4.91</td>
<td>Monitoring health behaviors in daily lives</td>
<td>6</td>
</tr>
<tr>
<td>Data sharing with health care providers</td>
<td>4.42</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Conclusion
The findings indicate that self-monitoring and tracking health indicators are the most difficult SM behaviors among people with chronic metabolic conditions. We conclude that the findings will contribute to designing and developing a novel mHealth tool to support sustainable health management of adults with metabolic conditions.

References
Towards Identification of Opiate Use Disorder in Clinical Text

John D. Osborne, PhD1, Tobias O’Leary, MS1, Abdullateef I. Almudaifer, MS1, Whitney L. Covington, MPH1, Caleb M. Carroll, BS1, Ellen F. Eaton, MD1, Rachel A. Lee, MD1, Sue S. Feldman, PhD1, Lauren Walter, MD1,

1University of Alabama at Birmingham, Birmingham, Alabama, United States of America

Introduction

The majority of overdose deaths in Jefferson County Alabama occurred in persons who received care in the University of Alabama at Birmingham (UAB) health system underscoring the many missed opportunities for Opiate Use Disorder (OUD) identification and treatment. Experts recommend universal screening for OUD in hospitalized patients in order to prevent adverse outcomes such as opioid withdrawal, elopement, infections, and overdose death. OUD requires a clinical diagnosis, yet patients may not disclose their disease due to stigma and criminalization. A prior retrospective study utilizing Natural Language Processing (NLP) revealed that over one-third of NLP-identified patients lack codes for opiate abuse or dependence1. We annotate and plan to release the first such public data set (to our knowledge) that could be used to identify OUD cases and expand capabilities for personalised intervention for OUD.

Method

Annotation was done using BRAT 1.3 software and documents were pre-annotated with an initial set of opiate keywords identified by regular expression. OUD researchers WC and CC were responsible for all annotation. Notes were double annotated for 2 development annotation rounds, revising annotation guidelines each round after discussion with TO and JO. 3295 notes from 59 patients (25 controls) were annotated. Annotators identified 42 label types including DSM-5 OUD symptom entities, comorbidities, treatments, drug tests, opiates and events. Annotators modified entities for negation, uncertainty, subject as assigned a DocTimeRel value of before, overlaps or after. Annotators identified explicit event mentions of Opiate Use, Substance Use, OUD and Substance Use Disorder (SUD) using the same set of modifiers as entities but with the addition of a severity and an illegal use modifier. Annotator agreement was calculated using Cohen’s Kappa in scikit-learn. Work was done under IRB-300002304 Initiating Medication Assisted Treatment of Opioid Addiction in the Emergency Department: The ED MAT Protocol funded by the Al. Dept. of Public Health.

Results

Annotator agreement was 0.702 and 0.965 in the 1st and 2nd annotation round respectively. The final data set consists of 3295 notes with 27953 mentions grouped under 12 higher level labels for brevity and shown in Table 1.

Table 1: Opiate Use Disorder Data Set

<table>
<thead>
<tr>
<th>Symptom Entities</th>
<th>Count</th>
<th>Other Entities</th>
<th>Count</th>
<th>Event Entities</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Loss of Control</td>
<td>509</td>
<td>Treatments</td>
<td>942</td>
<td>OUD</td>
<td>330</td>
</tr>
<tr>
<td>Pharmacological Problems</td>
<td>1531</td>
<td>Tests (Drug/HCV/HIV)</td>
<td>1866</td>
<td>SUD</td>
<td>180</td>
</tr>
<tr>
<td>Risky Use</td>
<td>737</td>
<td>Comorbidities</td>
<td>9015</td>
<td>Opiate Use</td>
<td>459</td>
</tr>
<tr>
<td>Social Problems</td>
<td>6</td>
<td>Substances (Drugs)</td>
<td>10488</td>
<td>Substance Use</td>
<td>1890</td>
</tr>
</tbody>
</table>

Conclusion

We demonstrate good annotator agreement and concept coverage, although social problems are poorly covered. We plan to release a de-identified set of notes and release an NLP-based algorithm for OUD case identification.

References

Usability Testing of MedWISER3: A Novel, Card-Based Electronic Health Record Interface
Soheb Osmani1, Katherine L. Dauber-Decker, PhD1, Sundas Khan, MD1, Sera Levy1, BA, Jennifer Itty, MPH1, Thomas McGinn, MD, MPH2,3, Jan Horsky, PhD1, Yalini Senathirajah, PhD2
1. Center for Health Innovations and Outcomes Research, Feinstein Institutes for Medical Research and Donald and Barbara Zucker School of Medicine at Hofstra/Northwell, Manhasset, NY; 2. Physician Enterprise, CommonSpirit Health, Chicago, IL; 3. iQuEST Institute, Baylor College of Medicine, Houston, TX 4. Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA

Description of the Problem
The gap between the way clinicians and programmers think often results in healthcare software that does not meet the clinicians’ needs. Clinician information requirements can vary widely by specialty, individual patient type, context, and institution.1 Current software interfaces are developed to have fixed information locations, which can impose undue extrinsic cognitive load.2 To address this problem, we sought to develop a novel, customizable, “draggable cards” based electronic health record (EHR) called MedWISER3.3 In this context, cards are: self-contained interactive blocks of information similar to widgets, comparable in size to playing cards, and can be dragged in the MedWISER3 interface to be relocated for customization.4 We performed usability testing to examine how clinicians obtain, utilize, and process information from MedWISER3. Additionally, we obtained feedback on the new EHR to optimize its utility.

Methods
We performed “think aloud” usability testing of the MedWISER3 interface with 11 internal medicine and endocrinology providers (two attending physicians, one hospitalist, two fellows, and six residents) at a large academic healthcare system in New York. Participants were asked to work through four mock patient cases on the MedWISER3 interface as they would with a real patient, providing a diagnosis and treatment plan at the end of each case. We performed thematic analysis of session notes and recordings to develop suggestions for improving MedWISER3.

Results
We identified three recurring themes in our usability testing sessions: functionality, content and usefulness, and visibility and navigation. The participants expressed that the functionality of MedWISER3’s card-based system could be further developed to be more refined and less clunky. The cards often overwhelmed the participants because of the vast quantity of information they immediately provided. The participants noted that the content and usefulness of MedWISER3 were improved compared to existing EHRs; however, the MedWISER3 interface lacked information standardization, which would allow rapid information processing. As expected, the participants expressed that the card-based EHR improved the overall visibility of content and the speed at which users could navigate through the EHR. Most participants expressed a strong interest in seeing reformed and improved EHRs.

Discussion
Several participants remarked on the patient notes in MedWISER3, reporting that it was difficult to follow the patient’s medical history. Furthermore, participants noted that using an EHR in which multiple patients’ charts could be open at the same time could cause confusion. A rework of the note tab would be most beneficial to the overall usability of the MedWISER3 interface. Our findings highlight clinicians’ need for a modernized EHR and how a card-based interface could streamline information processing. Importantly, several participants noted that this interface decreased the total time spent in the EHR. Reducing time and number of clicks in the EHR is crucial for providers who are moving quickly from patient to patient. Our future work will involve comparing MedWISER3 to an existing EHR.

Conclusion
Our observations highlighted several important areas in which EHR usability can be optimized for clinician use. There is a clear need to modernize EHR interfaces. MedWISER3’s card-based software could help bridge the gap between the clinician’s needs and the programmer’s development due to its layout and customizability.

References
Methodologies of Nursing Workflow Studies: A Scoping Review
Yennuten Paarima, MPhil, RN, Po-Yin Yen, PhD, RN, FAMIA, FAAN
Goldfarb School of Nursing, Barnes-Jewish College, BJC HealthCare & Washington University in St. Louis, St. Louis, MO

Introduction
Workflow is defined as “the set of tasks grouped chronologically or arranged sequentially into processes” that describes the procedural and sequential aspects of processes and offers a methodical narration of how work is conducted. Since the adoption of electronic health records (EHR), it has led to essential nursing workflow changes which may potentially affect patient care and safety. This has necessitated a critical need to examine the approaches employed in workflow studies. As a result, we conducted a scoping review to investigate how nursing workflow studies were conducted and interpreted and identify knowledge gaps that could inform future workflow studies examining the impact of the EHR.

Methods
We conducted searches in PubMed, Embase, CINAHL, Scopus, and Cochrane Library to retrieve articles published between 2005 and 2020. We used keywords such as “electronic health records” OR “electronic medical records” OR “computerized medical record”, AND “nursing workflow OR work process”, OR “task performance”, OR “time and motion studies”. We included studies with an explicit goal of examining the impact of the EHR on nursing workflow. We excluded non-English studies, editorials, and literature reviews.

Results
We identified 2289 articles in the initial stage. After de-duplication, a total of 15 studies were included. Disagreements for exclusion were resolved through discussion. Among the 15 studies, we found that twelve used time and motion (T&M) design, two used mixed-method design, and one used qualitative design. These designs allowed the continuous observation and analysis of the movement of nurses as they perform clinical tasks with particular attention to the amount of time required to perform each task. Fourteen studies used clinician-centered approaches to identify role-specific clinical activities, and one used patient-centered which focuses on services arranged around patient needs. The T&M approach was used to measure nurses’ time spend with the EHR in comparing to their time spent on performing other clinical tasks. We found that T&M studies measured the time nurses spend on both direct and indirect care activities. Nurses spend 18%-27% of their time charting and documenting in the EHR. However, we did not find any study investigating or reporting the changes in the sequential flow of nurses’ tasks which is suggestive that the impact of EHRs on nursing workflow has been mostly focusing on time efficiency.

Discussion
Several approaches as identified in this review have been used to assess how nurses distribute patient care activities with the EHR. The clinician-oriented approach focuses on a single nurse’s work instead of collaborative work, which is essential in healthcare delivery. However, observation of role-specific clinical activities will not likely represent the patient care provided. On the other hand, the patient-oriented approach captures various roles of the healthcare team in care delivery which promotes coordination and defining more meaningful care boundaries. Though workflow is defined as a set of tasks chronologically or the processes of care, we did not find any study investigating nursing workflow relative to the sequential order of tasks. Understanding the sequential order of nurses’ tasks will allow for proper streamlining and prioritizing of nurses’ activities and minimize interruptions. Nurses are spending more time with the EHR, but currently, there is no evidence of whether the time spent with EHR would impact other patient care activities. Understanding EHR activities along with patient care activities outside of the EHR may be the key to complete the whole picture of the nursing workflow. This will also allow for the investigation of nurses’ work fragmentation, multitasking, and interruption due to the EHR.

Conclusion
Existing workflow studies focused on nurses’ time spent in the EHR and did not directly address workflow changes. Future studies should consider using mixed approaches with data triangulation to explore the extent to which EHR can help restructure nursing workflow.

References
Comparison of a patient cohort and predictive models derived from local academic medical centers versus a national health database

Courtney Page1 MS, Conrad Sweitek2 BS, Cliona Molony2 PhD, Karen Chandross2* PhD, Benjamin Goldstein1, PhD
1. Duke Clinical Research Institute, Duke University, Durham, NC, USA
2. Sanofi, Cambridge, MA and *Bridgewater, NJ, USA

INTRODUCTION
The proliferation of real-world health data (RWD) has provided opportunities to understand people’s natural health evolution and care seeking behavior. A typical form of RWD for health is electronic health records (EHR) data, which contain detailed information on a patient’s health encounters. RWD sources differ in their underlying characteristics creating distinct advantages and disadvantages depending on the health question. Differences include size, geographic coverage and availability of source data.

Given these differences, an important uncertainty is the extent to which findings from one RWD source can be transported or replicated in another RWD source. In this project we sought to examine this question within the context of developing a predictive model for hospital admission for individuals with Type II diabetes.

METHODOLOGICAL APPROACH
Data Sources. Data were derived from Duke University Health System (DUHS); a large academic medical center that serves as the primary provider in Durham County, North Carolina, and a large national health plan that organizes EHR data and makes it available for research.

Data Extracted. We first defined a population of patients having Type II diabetes. We extracted information on demographics, service utilization history, comorbidities, procedures, medications, vital signs and lab test results.

Analytic Approach. We created Table 1 and compared the defined cohort using standardized mean differences (SMDs). For each data source, we used LASSO regression to derive a predictive model for hospitalization. We cross compared the model fit in independent test datasets.

RESULTS
We identified 18,355 diabetic patients from DUHS and 967,499 diabetic patients from the national data source. There were meaningful differences (most SMDs > 10%, with many > 20%) in the demographic and clinical characteristics of the patient groups. Of note, the DUHS cohort was younger and contained more females and African Americans than the national cohort. After limiting the national data set to just those living in the South, these differences attenuated but still persisted with many SMDs still greater than 10%, suggesting that differences were not solely due to geographic differences.

Predictive model results are shown in table below.

<table>
<thead>
<tr>
<th>AUC Results</th>
<th>DUHS in DUHS</th>
<th>DUHS in National</th>
<th>National in National</th>
<th>National in DUHS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>0.81</td>
<td>0.70</td>
<td>0.74</td>
<td>0.57</td>
</tr>
<tr>
<td>South Region</td>
<td>NA</td>
<td>0.59</td>
<td>0.65</td>
<td>NA</td>
</tr>
<tr>
<td>Caucasians</td>
<td>0.79</td>
<td>0.70</td>
<td>0.74</td>
<td>0.56</td>
</tr>
<tr>
<td>African Americans</td>
<td>0.82</td>
<td>0.69</td>
<td>0.69</td>
<td>0.58</td>
</tr>
</tbody>
</table>

DISCUSSION
While the national data are larger and arguably more nationally representative, the DUHS data have clearer data provenance and follow the PCORnet CDM. The DUHS data appeared more clinically accurate and ultimately produced a better performing predictive model that showed more transportability. The results suggest that a relatively well performing model built from a single center performed moderately well in a national dataset. Conversely, the nationally derived model did not perform as well in this same single center.
Applying NLP to extract clinician recorded affect to advance depression clinical research: A proof of concept

Vanessa Panaite, Ph.D., Dezon Finch, Ph.D., Andrew R. Devendorf, B.A., Lina Bouayad, Ph.D., Stephen L. Luther, Ph.D., MA, Susan K. Schultz, M.D. Research & Development Service, James A. Haley Veterans’ Hospital, Tampa, FL; Department of Psychology, University of South Florida, Tampa, FL; College of Public Health, University of South Florida, Tampa, FL

Problem addressed: Affective theories of depression suggest that affect marks the depressed state and is associated with depression severity, course, and outcomes [1]. Clinical practice is still in need of clinical markers that better capture patient desired outcomes. Current practice guidelines for depression use a symptom-management approach to gauging recovery [2] while patients place more value on achieving psychological well-being [3]. Affective flexibility has been linked to psychological well-being [4]. In the current pilot we proposed that clinician documented patient affect is reliably observed during mental health encounters and reliably extracted using natural language processing (NLP) to assist in the future with tracking patient desired depression outcomes over time. We piloted an information extraction vocabulary to test the feasibility and reliability of identifying clinician recorded patient affective states in clinical notes from electronic health records (EHRs).

Methods: Affect and mood were annotated in 149 clinical notes write by mental health providers for 109 patients. Annotations were completed by two independent coders. Affect was defined as a session specific emotional state that could be defined by emotion related behaviors and expressed or clinician observed emotional experiences. Mood was defined as affective behaviors and experiences over weeks or longer outside of session. Agreement statistics and frequency of the concepts were examined. Intercoder disagreements were settled by a third clinical coder. This reference annotation set was used to test a proof of concept NLP system using a Named Entity Recognition approach.

Results: Annotated data demonstrated that affect characteristics were identified in 88% of the notes; mood was identified in 97% of the notes. Reliability of affect and mood annotation was established across three coding waves. The intercoder reliability was consistently good across the three pilot coding waves (Inter-Annotator Agreement (IAA) = >70%) showing substantial agreement. The final NLP system showed good reliability with the final reference annotation set: affect IAA = 80.9%; mood IAA = 85.8%.

Conclusion: Concepts describing in session affect and general mood were frequently addressed in templated format and free text in clinical notes both by psychiatrists and psychologists. These findings show that concepts of affect and mood can be reliably identified in clinician reports and that they are good targets for NLP. This work has multiple implications: 1) it provides a solid springboard for a larger scope NLP to capture patient affect relevant to mental health treatment across multiple session; 2) supports the effort to better capture patient desired outcomes such as psychological well-being.

References:
Predicting Breast Cancer Treatment Outcomes for Minority Women in California Using Machine Learning

Jung In Park, PhD, RN¹, Doyub Kim, PhD², Kai Zheng, PhD³, Sunmin Lee, ScD, MPH⁴
Alpesh Amin, MD, MBA⁵

¹Sue & Bill Gross School of Nursing, University of California, Irvine, CA; ²NVIDIA, Santa Clara, CA; ³Donald Bren School of Information & Computer Sciences, University of California, Irvine, CA; ⁴University of California-Irvine, Department of Epidemiology, School of Medicine, Irvine, CA; ⁵School of Medicine, University of California, Irvine, CA

Breast cancer is the second leading cause of cancer-related deaths of women aged 40-55 in the United States. The National Cancer Institute (NCI) estimated that about 276,480 women would be diagnosed with breast cancer in 2020, and about 42,170 will die of the disease. Although white women have higher cancer incidents, minority women have a higher mortality rate. California has the largest ethnic/racial minority population in the United States – about 20% of the total national minorities live in California. Several studies have compared the breast cancer treatment outcomes, but few studies predict the treatment outcomes for minority women residing in California using large datasets. A large dataset allows robust modeling for individualized treatment decisions. This study aimed to build prediction models for breast cancer treatment outcome, which was 5-year survival, for minority women residing in California using a machine learning approach. The minority populations included in this study were Hispanic, Asian or Pacific Islander, Black, American Indian/Alaska Native.

We used the data from the Surveillance, Epidemiology, and End Results (SEER) program, which the NCI supports. SEER collects cancer incidence data from the cancer registries covering nearly 35% of the U.S. population. The data include patient demographics, primary tumor site, tumor morphology, stage at diagnosis, and the first line of treatment, and patients’ vital status. We developed prediction models using random forest (RF), XG Boost, SVM, and Logistic Regression, and these models were evaluated using precision, recall, and Area Under the ROC Curve (AUC). The grid search algorithm was used with 10-fold cross-validation to find the optimal hyperparameter set.

A total of 119,392 minority women who had breast cancer diagnoses between 2000 – 2017 were identified. After data preprocessing, the number of features used for modeling was 582. Most of them were Latino (48.1%), Asian or Pacific Islander (35.5%), Black (17.0%), and American Indian/Alaska Native (1.2%). The RF model found that HER2 positive, surgical primary site, summary cancer stage, whether surgery was performed, median household income, regional nodes positive were the predictors of the outcome. The most patients had localized stage (59.7%) followed by regional (35.5%), and Grade II (41.7%) followed by grade III (39.2%). The precision, recall, and AUC for each model were RF (87%, 69%, .86, respectively), XG Boost (87%, 70%, .86), SVM (86%, 70%, .85), and LR (86%, 71%, 0.85). Based on the evaluation results, the XG Boost model slightly outperformed the other models.

This study has developed highly accurate models to predict the breast cancer survival of minority women in California using large datasets and machine learning. It can work as an evidence-based decision support tool that is tailored to individuals for better patient outcomes. Future work is required to validate this algorithm in other health care systems to test its generalizability.
Technology and Data Sharing Preferences in mHealth Research Interventions
Leslie Park, BS1, Sabrina Mangal, PhD, RN1, Meghan Reading Turchioe, PhD, MPH, RN1, Annie C. Myers, MA1, Brittany Taylor, BS, RN2, Ruth M. Masterson Creber, PhD, MSc, RN1

1 Dept. of Population Health Sciences, Health Informatics Div., Weill Cornell Medicine, New York, NY; 2Columbia University School of Nursing, New York, NY

Introduction
The ubiquity of mobile devices enables an unprecedented level of accessibility to information, which has created an opportunity for mHealth interventions to empower patients to understand their health status and improve communication between patients and providers. As mHealth interventions grow, there are critical challenges in data governance, transparency, and accessibility that impact patients’ agency over health decisions and vulnerability to data misuse. The objectives of this study were: (1) to explore Heart Failure patients’ technology experiences in a mHealth intervention and (2) to gauge public trust and data sharing preferences regarding clinical research studies.

Methods
This is a two-part mixed methods study. Part 1 included qualitative interviews with patients with heart failure who had participated in an mHealth intervention that evaluated patient symptoms over eight weeks. These interviews were audio-recorded, transcribed, and deductively coded by a team of three trained researchers. Part 2 was a survey that focuses on a representative U.S. population’s technology experiences and attitudes about data access and sharing. The survey includes focused questions that were generated from our interview themes about attitudes towards technology use, data sharing, and trust in various institutions involved in delivering mHealth interventions.

Results
Part 1) Interviews with heart failure patients (n=16; average age: 63.6 years old, 81% male, 75% White, 19% Hispanic/Latino) participating in our mHealth intervention generated four emerging themes and illustrative quotes:

<table>
<thead>
<tr>
<th>1) Mobile devices make research participation convenient</th>
<th>“The phone makes it easy because it’s with me and I can do it whenever I want [...] I don’t have to worry about where I am or what type of day it is.”</th>
</tr>
</thead>
<tbody>
<tr>
<td>2) Email is a preferred mode of communication and record keeping for health information, but is often lost</td>
<td>“Email [is preferred], I probably wouldn’t go on any website because it’s difficult to track passwords. If I had a PDF and an email, I would dump it in a folder and leave it there. I have 10,000 emails from [years ago]”</td>
</tr>
<tr>
<td>3) Barriers to data sharing include concern about provider-burden and time constraints</td>
<td>“If [my provider] is okay with receiving [my symptom results], I would gladly share it, but I know she has a lot of patients. I don’t want to be a burden. When I do call her, I want an answer.”</td>
</tr>
<tr>
<td>4) Altruistic motivations for sharing their data with providers, researchers, and tech companies</td>
<td>“I figured [participating and sharing data] couldn’t hurt. It could only help if someone else is going through what I’m going through. It’s pretty scary going through it and not knowing what the outcome is going to be.”</td>
</tr>
</tbody>
</table>

Part 2) Survey (n=502; average age: 46.6 years old; 49% male; 74% white; 7% Hispanic of Latino) results revealed three key findings: 1) the internet, tech companies, large corporations, and government are significantly less trustworthy than health policy makers and healthcare providers; 2) mixed feelings of trust in sharing data with health tech companies for mHealth interventions (15.9% decrease trust, 52.4% no change in trust, 31.7% increase trust) with 36% preferring to share none of their health information with health tech companies in context of an mHealth study; 3) 72% felt that they should be paid for sharing data with health tech companies.

Conclusion and Discussion
This study raises important design and ethical considerations for implementing mHealth interventions, especially those involving tech companies either directly as partners or indirectly as intermediaries that provide the devices, applications, data, and systems infrastructure necessary for implementing scalable solutions involving mHealth. Future mHealth studies should consider implementation design choices, protocols for supporting information accessibility and actionability, patient-provider communication, as well as equitable and ethical data access and use.

[Study support: National Institute of Nursing Research (R00NR016275; R00NR01675-05S1; PI: Masterson Creber)]
Assessing the Predictive and Ecologic Validity of Geographic Indices for Social Determinants of Health
Yoonyoung Park, ScD1, Frank Yoon, PhD2, William Kassler, MD, MPH2
1IBM Research, Cambridge, MA; 2IBM Watson Health, Cambridge, MA

Introduction The growing interest in social determinants of health (SDoH) that affect public health and well-being has led to development of geographic deprivation indices (GDIs) that characterize SDoH based on geography. While most component measures are sourced from population statistics, GDIs largely differ in concepts and analytic methodology. This work is motivated by lack of evidence on the comparative performance of popular GDIs to inform analytic choices for health policy research and practice. We evaluate the predictive performance of four existing GDIs and examine the threat of ecological bias to inform the choice of GDI in practice.

Methods The four candidate GDIs examined in this study are: Area Deprivation Index (ADI), Community Need Index (CNI), Social Deprivation Index (SDI), and Social Vulnerability Index (SVI). Conceptually, the ADI captures health inequity, the CNI the demand for health services, the SDI the relative levels of social disadvantage, and the SVI the vulnerability and preparedness for natural disasters. Except for SVI which is based on percentile ranks, the three other GDIs were developed using factor analysis. All four indices are available at county level. The area under the receiver operating characteristic curve (AUROC) was calculated by logistic regression models predicting the worst decile of life expectancy, diabetes prevalence, and preventable hospital admissions. To assess ecological bias, we obtained health measures from County Health Rankings & Roadmaps (county-level data) and IBM Watson Health™ PULSE® Healthcare Survey and Claritas PRIZM® Premiere Segmentation (block group-level data) for the state of New York. Using generalized linear models, we incorporated the ADI at different geographic levels in predictive models for emergency room (ER) utilization to assess at what geographic level the ADI becomes less informative or changes the direction of effect, with or without adjusting for other SDoH measures (18 chronic condition indicators, insurance type, and SDoH such as income level, household composition, home ownership, education, and employment).

Results The GDIs exhibited strong pairwise correlations (0.73-0.86). In predictive models, ADI consistently showed the best performance in terms of AUROC values (Table 1). Adding ADI to the models adjusting for age categories and subsequently additional variables, we observed that the signs of ADI coefficients change when block-group or tract level ADI is used compared to when county level ADI is used (Table 2). The magnitude of ADI coefficients generally decreased as the aggregation level increased or as additional SDoH variables were added.

Table 1. AUROC of county-level GDIs

<table>
<thead>
<tr>
<th></th>
<th>Bottom 10% life expectancy</th>
<th>Top 10% Diabetes prevalence</th>
<th>Top 10% Preventable hospital stays</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADI</td>
<td>0.904</td>
<td>0.875</td>
<td>0.792</td>
</tr>
<tr>
<td>CNI</td>
<td>0.782</td>
<td>0.752</td>
<td>0.651</td>
</tr>
<tr>
<td>SDI</td>
<td>0.850</td>
<td>0.801</td>
<td>0.693</td>
</tr>
<tr>
<td>SVI</td>
<td>0.835</td>
<td>0.798</td>
<td>0.692</td>
</tr>
</tbody>
</table>

Discussion We observed differences in predictive performance across four GDIs based on AUROC. In general, ADI performed well and reliably across experiments. But the impact on performance in practice, such as in risk adjustment models, may not be as strong if GDIs are used as one of many features in a model. The differences in concepts, components, and associated face validity should be considered when choosing a GDI for an analytic task. One important factor to consider is the aggregation level, as we can see some evidence for the presence of ecological bias when county-level data is used in place of individual or more granular data. We suggest that GDI should be used at the highest geographic resolution whenever possible to avoid such limitations.

References
3. Butler DC, Petterson S, Phillips RL, Bazemore AW. Measures of social deprivation that predict health care access and need within a rational area of primary care service delivery. Health Serv Res. 2013 Apr;48(2 Pt 1):539-59

Table 2. Linear regression model coefficients for ADI variable

<table>
<thead>
<tr>
<th></th>
<th>Block group-ADI</th>
<th>Tract-ADI</th>
<th>County-ADI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 1 (Age)</td>
<td>0.2194</td>
<td>0.2088</td>
<td>0.0604</td>
</tr>
<tr>
<td>Level 2 (+ chronic conditions)</td>
<td>0.0788</td>
<td>0.0667</td>
<td>-0.0115</td>
</tr>
<tr>
<td>Level 3 (+ Insurance)</td>
<td>0.0471</td>
<td>0.0337</td>
<td>-0.0265</td>
</tr>
<tr>
<td>Level 4 (+ SDoH)</td>
<td>0.0269</td>
<td>0.0172</td>
<td>-0.0287</td>
</tr>
</tbody>
</table>
Use of a Widely Available Electronic Health Record to Facilitate COVID-19 Vaccinations

Mark A. Parkulo, MD 1,6; Windell Smith, MHA 2, Luanne Williams, MSN RN NE-BC 3, Marites E. Sayo, MBA 2, Jason Buckmeier 4, Jodi Rinken 4, John C. O’Horo, MD, MPH 5

1Mayo Clinic, Jacksonville, FL: Division of Community Internal Medicine, Department of Medicine, 2 Department of Administration, 3 Department of Nursing, 4 Mayo Clinic, Rochester, MN: Department of Information Technology, 5 Division of Infectious Diseases, Department of Medicine, 6 Mayo Clinic Robert D. and Patricia E. Kern Center for the Science of Health Care Delivery.

Background  On December 23, 2020 Ron DeSantis, governor of the state Florida, issued an executive order authorizing COVID-19 vaccination of healthcare workers, patients over the age of 65 and other patients who were extremely vulnerable. A process to prioritize the highest risk patients was implemented at Mayo Clinic in Florida. 1,2

Methods  Initial evaluation of patient data generated by our Electronic Health Record (EHR) (Epic Systems, Inc.) revealed 182,265 unique patients from the state of Florida had at least 1 visit/encounter in 2020 with 61,488 of those patients being over age 65 and living in the surrounding 5 counties. Utilizing a COVID Risk Score3 (based on demographic, medication and comorbid risk factors) to predict risk of severe complications of COVID19 infection, we were able to further segment the population and invite highest risk patients first.

Results  Between January 4 and January 7, 2021, we invited 11,000 patients (Table 1) and were rapidly able to fill 4,093 scheduled appointments with 3,156 patients able to self-schedule online or via the Mayo Clinic app. (Table 2). Between January 5 and January 14, 2021, we administered 4,093 scheduled vaccines to patients until our vaccine supply was exhausted. We were able to rapidly fill schedules utilizing this process and placed invited patients requesting vaccine on a waitlist when appointment calendars were filled (Table 1). We halted the process when our vaccine supply was exhausted. We reinitiated the process when more vaccine was delivered to us by the State.

Table 1. Characteristics of Patients Invited and Accepting Invitation as of Jan. 25, 2021

<table>
<thead>
<tr>
<th>Invitations</th>
<th>Total</th>
<th>Median Risk Score</th>
<th>Avg Risk Score</th>
<th>Median Pt Age</th>
<th>Avg Pt Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>Portal/Mail</td>
<td>8963/2037</td>
<td>5</td>
<td>5.28</td>
<td>77</td>
<td>78.05</td>
</tr>
<tr>
<td>Administered/Waitlist</td>
<td>4093/697</td>
<td>5</td>
<td>5.3</td>
<td>76</td>
<td>77.1</td>
</tr>
</tbody>
</table>

Table 2. Scheduling and Administration Timing in Days

<table>
<thead>
<tr>
<th></th>
<th>Self-Scheduled</th>
<th>Staff Scheduled</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median Order to Schedule Date</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Avg. Order to Schedule Date</td>
<td>0.6</td>
<td>1.9</td>
</tr>
<tr>
<td>Median Schedule to Appointment Date</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Avg. Schedule to Appointment Date</td>
<td>2.9</td>
<td>2.4</td>
</tr>
<tr>
<td>Median Order to Vaccination Date</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Avg. Order to Vaccination Date</td>
<td>3.5</td>
<td>4.3</td>
</tr>
</tbody>
</table>

Conclusion  EHRs can facilitate rapid identification and prioritization of high-risk patients, efficiently communicate eligibility, and enhance scheduling and administration processes for COVID 19 vaccinations. 4 The utilization of electronic communication and scheduling techniques allowed administrative personnel to focus on patients who needed manual assistance.

References

Core Outcomes Sets Development in COVID-19 Clinical Trials
Irena Parvanova and Joseph Finkelstein
Icahn School of Medicine at Mount Sinai, New York, NY, USA

1. Introduction
A novel coronavirus disease (COVID-19) has caused a worldwide pandemic outbreak [1]. As of January 2021, the virus affected the population of 222 countries, with 83,910,356 confirmed cases at the time. Outcome reporting and standardization in COVID-19 clinical research depends on the use of a core outcome set (COS), collected and compared across all COVID-19 trials [2]. The growing research data are currently being integrated in data repositories, including publicly available repositories, such as ClinicalTrials.gov.

2. Methods
Primary and secondary outcomes from COVID-19 trials in ClinicalTrials.gov were analyzed using analytical query, combining “COVID-19” in the “Other terms” field with “United States” in “Country” and “NIH” in “Funder Type” fields, generating 120 studies at the time of our inquiry. We performed detailed analysis where we examined common data elements based on the studies outcomes.

3. Results
We examined the outcomes for the 120 clinical trials and ranked each outcome for the COVID-19 clinical trials by calculating the percentage of the appearance of each outcome in all 120 trials. For quantification purposes, we considered outcomes that were present in more than 1% of the studies. The most common COVID-19 clinical outcomes are: Mortality due to COVID-19 (16%), Mental health impact due to COVID-19 (12%), Anti-SARS-CoV2 Antibodies (10%), Proportion of patients that required hospitalization due to COVID-19 or died (8%), SARS-CoV-2 PCR/RT-PCR test results/nasal swabs samples (6%), Severity of COVID-19 (6%), and others (Figure 1).

4. Discussion
ClinicalTrials.gov contains enormous quantity registered clinical trials information, including COVID-19-related clinical trials. Our findings provide standardized approaches for representation of COVID-19 research to allow for cross-trial comparison of the results and building a comprehensive core outcome sets in COVID-19 clinical trials. Synchronization of common data elements and data exchange standards is necessary to facilitate data harmonization approaches in COVID-19 research.

5. Conclusion and future work
Future plans include implementing ontology-driven automated metadata extraction pipeline from ClinicalTrials.gov, combined with crowdsourcing for engaging multidisciplinary researchers in ranking COS candidates [3].

References
Optimizing Clinician Alerts for Concerning Symptoms in an Electronic Patient-Reported Outcomes (ePRO) System in Community Oncology

Amila M. Patel, PharmD¹, Lalan S. Wilfong, MD², Ethan Basch, MD, MSC³, Angela M. Stover, PhD⁴ Max Franklin¹, Ben Pearson¹, Debra A. Patt, MD, PhD, MBA²

¹Navigating Cancer, Seattle, Washington; ²Texas Oncology, Austin, Texas; ³University of North Carolina at Chapel Hill, Chapel Hill, North Carolina

Introduction

Successful implementation of electronic patient-reported outcomes (ePROs) tools in cancer care is highly dependent on integration with routine clinic workflows for symptom management. This includes appropriate and timely notification to clinical staff for symptom intervention. However, there is a risk of clinician burden and both patient and provider dissatisfaction if ePRO symptom notification algorithms fail to prioritize symptoms warranting attention. Identifying patients who need acute symptom intervention may help optimize staffing and improve time to symptom resolution.

Methods

Texas Oncology, a large multi-site community oncology practice, partnered with Navigating Cancer to implement an ePRO technology across 210 sites of service for symptom monitoring among patients initiating a new systemic therapy. Patients were sent symptom assessments on a weekly basis to characterize 14 common cancer-related symptoms for presence, frequency, severity, and/or interference with daily activities based on NCI’s PRO-CTCAE instrument, with responses based on a 5-point Likert scale. Participating patients self-reported via email or text and moderate-severe symptoms triggered a real-time notification to nursing staff. We systematically tracked clinician alerts and interventions, and continuously collected user feedback. These data were reviewed by a multi-disciplinary stakeholder group and practice leadership to modify the alert algorithm based on consensus.

Results

6,867 patients completed 61,688 symptom assessments between July 2020 and January 2021. 68% of reported symptom items triggered an alert. 46% of alerts were associated with a single symptom, with the most common being fatigue in the absence of other symptoms (13%). These fatigue-only alerts resulted in no direct clinical intervention by nursing staff, but rather reinforcement of lifestyle and behavior modifications for fatigue management. A decision was made to remove fatigue-only alerts from the algorithm, but to continue providing patient-reported fatigue data to providers at clinic visits.

Discussion

A high volume of clinician alerts for concerning symptoms occurred, and thus we created a framework to optimize the alerting algorithm, with a specific focus on fatigue-related alerts. Although fatigue is a significant patient burden, it did not elicit immediate intervention by nursing staff, possibly because there are limited available acute interventions for fatigue. An algorithm prioritizing symptom alerts based on clinical actionability enables nurses to focus on symptoms that warrant acute intervention.

Conclusion

Systematic analysis and multi-disciplinary review of ePRO alerts and resulting nursing actions enables a deliberate approach to improving alert algorithms and optimization of clinician time. Further evaluation of the influence of predictive algorithms to tailor alerts to clinically significant criteria coupled with the use of automated tailored patient education materials on utilization, efficacy, and clinical staff feedback will be studied.
From Emergency Department to Admission: mapping reasons for visit and admit diagnosis using Natural Language Processing

Olga V Patterson, PhD1,2, Hannah Eyre, MS1,2,
Kelly S. Peterson 1,2, Scott L DuVall, PhD1,2
1VA Salt Lake City Health Care System; 2University of Utah, Salt Lake City, UT, USA

Introduction

Even with a broad spectrum of medical imaging and laboratory tests that provide accurate and objective information about a patient’s state, the importance of symptom-oriented research was highlighted during the COVID-19 pandemic. The Veterans Affairs (VA) electronic medical record system contains data collected at the point of care, which are aggregated nightly into the Corporate Data Warehouse and made available for research. Among a number of other data elements, the reasons for Emergency Department (ED) visits and the inpatient admit diagnoses are entries that indicate patients’ presenting symptoms or diagnoses. These entries are implemented as 50 character-fields with free text entries or automatically populated text that represents ICD10-CM codes if specified by the user.

Methods

In order to map free text entries to a standard vocabulary we developed a natural language processing (NLP) system* that contained the following modules: 1) for text preprocessing, we adapted patterns from the Emergency Medical Text Processor (EMT-P)1 by adding or removing some patterns to perform initial spelling correction and abbreviation expansion, 2) for UMLS concept mapping, we utilized cTAKES 4.0 Fast Dictionary Lookup module2 with UMLS 2020AA dictionary, and 3) for context detection, we utilized ConText algorithm3 expanded to VA specific terminology. The NLP pipeline was implemented within Leo infrastructure4 that is based on UIMA AS 2.9.0. The system was validated on 500 distinct presenting symptom statements through double annotation.

Results

Validation showed that the system achieved 91.4% precision and 83.4% recall in identifying mentioned concepts and their contexts. As of Dec 31, 2020 the VA COVID-19 Shared Data Resource5 had 449,117 patients who had 854,699 ED visits in 2020. The concept mapping NLP system was used to process 1,504,319 statements of ED visit reasons and 644,487 of inpatient admit diagnoses. The most frequently reported ED symptoms were dyspnea (11.7%), COVID-19 (8.1%), and chest pain (6.7%). The most frequently negated symptoms were homicidal/suicidal ideation, COVID-19, and dyspnea. In most cases, patients were released home (59.7%). Out of those who were admitted to the hospital, the most frequent admit diagnoses were COVID-19 (5.8%), pneumonia (4.2%), and chest pain (3.9%).

Discussion

Biosurveillance requires real-time recognition of emerging infections and symptom-based research depends on accurate detection of symptoms. We built a lightweight high-performing NLP system and deployed it in April 2020 for comprehensive tracking of presenting symptoms for the Veterans who were tested for COVID-19 within the VA system. Over 140 studies are currently in progress that utilize the chief complaints recorded in the VA EDs.

Conclusion

We developed a scalable, high-throughput system that provides concept mapping for patient visit reasons and inpatient admit diagnoses. The system was implemented in April 2020 and has been in use to provide data for the VA COVID-19 Shared Data Resource with weekly dataset updates.

Acknowledgement

This work was supported using resources and facilities of the VA Informatics and Computing Infrastructure (VINCI), VA HSR RES 13-457.

References


*https://github.com/department-of-veterans-affairs/ccmap
†https://www.hsrd.research.va.gov/for_researchers/cyber_seminars/archives/video_archive.cfm?SessionID=3810
Identifying erosive disease from radiology reports of veterans with inflammatory arthritis using natural language processing

Shaobo Pei, PhD1,2, Gopi K. Penmetsa, MD1,2, Brian C. Sauer, PhD1,2, Bingjian Feng, PhD3, Kevin Douglas, MD3, Jerry Clewell, PharmD3, Jodi Walker, PharmD3, Jessica A. Walsh, MD1,2
1George E. Wahlen Department of Veterans Affairs Medical Center, Salt Lake City, Utah; 2University of Utah, Salt Lake City, Utah; 3AbbVie, Inc., North Chicago, Illinois

Introduction
The presence of erosive disease influences diagnosis, management, and prognosis in inflammatory arthritis (IA). Observational research of IA in large datasets has been limited by a lack of methods for identifying erosions. We developed methods for identifying articular erosions in radiology reports from veterans with IA.

Methods
Included veterans had ≥2 ICD9 or ICD10 codes for ankylosing spondylitis (AS), psoriatic arthritis (PsA), or rheumatoid arthritis (RA) between 1/1/2005-12/31/2019, in the Veterans Affairs Corporate Data Warehouse. Chart review and annotation of relevant radiology notes produced the reference standard, and identified erosion terms that informed classification rule development. A rule-based natural language processing (NLP)1 model was created and revised in training snippets. The NLP method was validated in an independent reference sample of IA patients at the snippet (sections of text containing 30 words before and 30 words after the meaningful term) level and patient level.

Results
In 168,667 veterans with IA, the mean age was 63.1 and 90.3% were male. These IA patients involved 2,926,113 radiology notes with 35,141 note titles. Based on 11 root terms (ero*, pencil*cup, cort*, irreg*, etc.), we found 1178 variations (erosion, erotic, etc.), and selected 179 possible terms (erosion, erode, etc.). Within 5000 randomly selected erosion snippets, we identified 6 meaningful terms representing erosion (pencil* cup, erosion, erosive, etc.), 28 terms representing erosive process irrelevant to joint(s) (i.e. gastric erosion), 5 terms representing non-IA processes (i.e. infection, trauma). Based on annotated data, 83 rules were developed for assigning the target concept “erosion” using the surrounding contextual language (i.e. classifying as negated vs. affirmed, related vs. unrelated to a joint or bone, due to vs. not due to IA). NLP models on snippet level or patient level were built and optimized on training dataset with 817 snippets (AS 417, RA 200, PsA 200) and testing dataset with 200 snippets (AS 100, RA 50, PsA 50). The models were validated on independent dataset of 60 randomly selected IA patients (20 AS, 20 RA, 20 PsA) with 254 snippets (53 AS, 122 RA, 79 PsA). The accuracy of the NLP model was 96% at the snippet level and 93% at the patient level, for all IA patients (Table 1).

<table>
<thead>
<tr>
<th>Training (n=817 snippets, 39% affirmed cases)</th>
<th>Testing (n=200 snippets, 36% affirmed cases)</th>
<th>Validation (n=254 snippets, 40% affirmed cases)</th>
<th>Validation (n=60 patients, 47% affirmed cases)</th>
</tr>
</thead>
<tbody>
<tr>
<td>IA</td>
<td>AS</td>
<td>RA</td>
<td>PsA</td>
</tr>
<tr>
<td>Recall</td>
<td>0.95</td>
<td>0.90</td>
<td>0.99</td>
</tr>
<tr>
<td>Precision</td>
<td>0.96</td>
<td>0.96</td>
<td>0.96</td>
</tr>
<tr>
<td>F1-score</td>
<td>0.95</td>
<td>0.93</td>
<td>0.97</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.97</td>
<td>0.95</td>
<td>0.98</td>
</tr>
</tbody>
</table>

Table 1. Performance of methods for identifying erosive disease. Inflammatory arthritis (IA) = ankylosing spondylitis (AS), rheumatoid arthritis (RA), and psoriatic arthritis (PsA).

Conclusion
The methods accurately identify erosions documented in radiology reports of veterans with IA. They may facilitate a broad range of research involving cohort identification, disease severity stratification, diagnosis, and treatment.

References
Modernization of the Intermacs Data Warehouse: Moving Data Management from SAS to SQL Server for Analytics

John K. Pennington, MSHI¹, Nicholas Timkovich, MSHI¹, Bunyamin Ozaydin, PhD¹
¹University of Alabama at Birmingham, Birmingham, AL, USA

Introduction

The Kirklin Institute for Research in Surgical Outcomes (KIRSO) mechanically assisted circulatory support device (MCSD) registry named Intermacs utilizes monthly flat-file data extraction from Microsoft SQL Server. Extraction is executed through a web interface triggering a series of SQL Server Views denormalizing stored data, and culminating resulting records in text files containing a comprehensive snapshot of data. A series of SAS Analytics Software programs are then executed against those text files rendering formatted and labeled SAS datasets for use in statistical analysis and outcomes reporting to contributing hospitals for benchmarking. This poster will present a rework of the extract, transform, and load (ETL) process such that data are not extracted from SQL Server as flat text files and manipulated by SAS. Instead, information within SQL Server is transcribed into a set of denormalized SQL tables structured to match those resulting from the existing SAS datasets. This workflow transformation ensures the ETL and data management processes are executed wholly within SQL so that KIRSO sees operational benefits via streamlining code management and personnel responsibilities, and creates a path towards further optimization and implementation of additional features through automation and business intelligence software integration¹.

Methods

The data warehouse utilizes a sample of over 25,100 adult patients receiving mechanical assisted circulatory device (MCSD) treatment for advanced heart failure from one of 181 participating Intermacs hospitals within the United States. The ETL process redesign consists of four main phases: SQL DDL, SQL DML, SQL Data Processing, and a SAS Macro. The SQL DDL phase creates table shells for denormalized data extracted from SQL Server. The SQL DML phase moves those data from their storage locations on SQL Server, denormalizes them, and renames columns in accordance with SAS variables. The SQL Data Processing phase derives new columns necessary for reporting that are not stored or calculated via the web application, as well as making corrections to some of the problematic data manager created data records. The SAS Macro is a SAS program acting as its own mini-ETL by allowing SAS users to extract SQL tables, transform them into formatted and labeled SAS datasets, and load them either into SAS working memory or a network location by an Open Database Connectivity (ODBC) connection².

Results and Conclusion

The Intermacs data mart was validated as a successful process redesign by cross comparisons, including row and column counts as well as SAS proc tabulate of key dates and SAS proc freq of adverse event and outcomes variables, ensuring equilibrium between data created through the legacy SAS workflow with data created in the new SQL workflow. The data mart and data management redesign expedited data preparation, reducing the lag time between date entry and retrieval for research and made data accessible to a broader audience. The process redesign allowed for a simpler process when updating and creating research datasets by eliminating some of the personnel and existing manual interim steps in the ETL process that were artifacts of the legacy system. Legacy workflow averaged 32 minutes of SAS run time compared to an average of 17 minutes of SQL query time. This data processing time reduction of 47% becomes a 100% reduction in personnel resources via off-hours automated execution as nightly SQL stored procedure data refreshes. The data that were subsequently available also facilitated the future roll-out of timelier reporting as data are now stored in the KIRSO EDW on SQL Server and may be refreshed nightly instead of quarterly. This more frequent refresh narrows the reporting window from previous one to three month gap from data entry by a participant until those data were delivered in reports. Lastly the warehouse redesign also laid the groundwork for future real-time interactive reporting, where sites might request customized cohorts and exhibits through the web interface in addition to receiving exhibits based on cohorts as designed by the KIRSO Data Collection and Coordinating Center (DCCC).
Comparison of Population-wide Explanations for Predicting the Outcomes of Patients with Community-Acquired Pneumonia

Eddie Pérez, M.S.¹, Shyam Visweswaran, M.D., PhD¹, and Harry Hochheiser, PhD¹
¹University of Pittsburgh, School of Medicine, Department of Biomedical Informatics, Pittsburgh, PA

Introduction
Wider adoption of machine learning (ML) models in healthcare settings requires model explanations that are human interpretable, locally accurate, and globally consistent (1). Shapley Additive Explanations (SHAP) is a novel tool that theoretically guarantees the local accuracy of these explanations (2). To effectively guide patient care, explanations will ideally provide consistent insight into patient state, independent of the models used. However, the consistency of explanations across different types of models has thus far not been a topic of significant investigation.

Methods
In this study, we evaluate the consistency of explanations across popular ML models (Logistic Regression (LR), Random Forest (RF), and Gradient Boosted Machines (XGBM and LGBM)) when predicting clinical outcomes for community-acquired pneumonia (CAP) patients. Lundberg et al. showed that LR and GBM’s can weigh features differently resulting in differing explanations (1). These results were attributed, in part, to differences in each model’s bias/variance balance. Thus, hypothetically, the different models compared in this study could yield differing explanations.

The models were trained on a subset of the dataset collected by the Pneumonia Patient Outcomes Research Team (PORT) (3). Our pneumonia dataset contains 2,287 patients and has been studied extensively with the goal of training models to predict a dire outcome (i.e., serious complications, including death) of CAP (3). We used the following python libraries: [1] scikit-learn, xgboost, and lightgbm for training models, [2] SHAP for explanations, [3] Scipy to calculate Kendall’s W coefficients and p-values.

Results
Pairwise explanation consistency ranged from 0.45 to 0.78 (Table 1). Maximum consistency was 0.78 LGBM and XGB, both of which are GBM algorithms. RF had the minimum average consistency 0.47.

Conclusion
Model explanations differ across training algorithms. There was more agreement between LR and the GBMs than between RF and the GBM’s. Further research is necessary to understand how these differences affect the relevance of the model explanations to the clinical problem.

References
Early Prediction of Positive Clostridioides Difficile Test Results

Anh Pham, MS1, Robert El-Kareh, MD1, Lucila Ohno-Machado, MD, PhD1, Tsung-Ting Kuo, PhD1

1 University of California San Diego, La Jolla, CA, USA

Keywords. Machine Learning, Data Mining, Clinical Decision Support

Introduction. Clostridioides difficile infection (CDI) caused by the Clostridioides difficile (C. difficile) bacterium may lead to serious complications, especially in older, antibiotic-treated patients.1 While early detection of carriers could help manage outbreaks,2 current guidelines do not recommend broad screening at admission.3 A CDI risk-score model can identify at-risk patients for targeted testing. Past models are subject to issues of generalizability and feasibility.4,5 We sought to predict positive CDI test results (which may differ from true CDI diagnoses) with high accuracy, interpretable results, and a reasonable number of covariates, adding to the physicians’ decision-support toolbox with other testing strategies.

Materials and Methods. Our models predict C. difficile test results up to one calendar day before the first actual positive test, using 104 covariates of 157,493 UCSD Health patients from 01/01/2016 – 03/07/2019 (our IRB date). There were 1,541 cases which had at least one positive C. difficile lab test. Predictors include (a) race/age/gender, and (b) medication used in the timeframe, grouped by pharmacologic classes. We chose Logistic Regression (LR) and Random Forest (RF) algorithms, evaluated by the Area Under the Receiver Operating Characteristic Curve (ROC AUC). We held out 10% of the dataset for testing and tuned hyper-parameters via grid search using 10-fold cross validation on the remaining 90%. From selected hyper-parameters, the LR and RF models were retrained on 90% of data. Afterwards, an ensemble model averaged the LR and RF predictions to find the final scores and compared them to a threshold of 0.5 to decide outcomes. We performed significance analysis on the model covariates, recording their LR odds ratios and corresponding p-values, and RF Gini importance.

Results. The ROC AUCs are 0.839 (LR), 0.851 (RF), and 0.866 (ensemble). Among the significant factors (p < 0.0001), the clinically seen risk factors are age, use of anti-infectives,6 and gastrointestinal (GI) tract medications.

Conclusion. Our models using patients’ medication history and demographic can predict first positive CDI tests. For infectious diseases, early prediction of positives may aid clinical practices.

Acknowledgements. The authors were funded by the U.S. National Institutes of Health (NIH) (R00HG009680, R01HL136835, R01GM118609, R01HG011066, U24LM013755), a UCSD Academic Senate Research Grant (RG084150), and the Graduate Division San Diego Matching Fellowship associated with San Diego Biomedical Informatics Education & Research (SABER) NIH National Library of Medicine (NLM) grant T15LM011271. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The authors would like to thank Michael Hogarth, MD and UCSD ACTRI.

References

Integrating Molecular and Clinical Data for Real World Evidence Research in Precision Oncology

Scott Phillips, B.S.¹, Xiaojun Xu Ph.D.¹, Paulina Paul, M.S.², Wenhong Zhu PhD³, Michael Hogarth M.D.²,³, Olivier Harismendy, PhD ¹,²

¹Moores Cancer Center, ²Division of Biomedical Informatics, Department of Medicine, ³Altman Clinical and Translational Research Institute, University of California San Diego School of Medicine, CA

In precision oncology, Real World Evidence (RWE) research depends on clinicopathological data (stage, grade, diagnosis, outcome), which is not often available as structured data and hence requires manual curation of Electronic Health Record (EHR) data. Moreover, the practice of precision oncology increasingly relies on the results of DNA sequencing assays (NGS), which are available only as narrative text reports and not adequate for large-scale research use. Research initiatives such as the AACR GENIE consortium are aggregating data, relying on their members to share de-identified test reports accompanied with a minimal set of clinical information. Similarly, prospective clinical studies have collected molecular data and curated the clinical information to evaluate predictive biomarkers. Here we present an alternate approach that minimizes curation in the assembly of molecular and clinical data and which was used to create the Aggregated Registry for Molecular Oncology Research (ARMOR), a large-scale registry supporting RWE research in precision oncology.

Structured NGS report data were collected directly from the laboratories, extracted, and transformed to flat files following the cBioportal extended formats. The associated meta-data including laboratory, assay-type and version, diagnosis, specimen site and primary organ at order, Tumor Mutational Burden, or Microsatellite instability are added to sample and patient tables. The clinical and demographic data is obtained from the UCSD Health Clinical Data Warehouse for Research and included antineoplastic prescriptions and start dates, which were abstracted to their main ingredient using the Anatomical Therapeutic Chemical (ATC) classification. The data is anonymized, and all dates converted to duration since first report and made available to approved investigators via a local cBioportal server.

A total of 6,098 DNA sequencing test reports from 4,900 patients were collected between 2012 and 2020. The distribution of cancer types was consistent with the prevalence of cancer diagnosed at more advanced stages and more likely to be subjected to molecular testing (Lung: 10%, Colon: 9%, Breast: 7%, Prostate: 5%). We identified medication prescriptions and their dates for a subset of patients (N=3,970). A total of 360 antineoplastic medications, mapped to 145 ingredients. A total of 64 ingredients were classified as targeted therapy (TT), including 40 protein kinase inhibitors and 24 monoclonal antibodies. The usage of antineoplastic compounds across cancer type was consistent with the disease specific treatment (e.g. temozolomide in brain cancer) and highlighted frequent combinations (e.g. fluorouracil, irinotecan and oxaliplatin in colorectal cancer). The majority of the molecular tests were gene panel sequencing from tumor DNA (5,090, 83%) or cell free DNA (971, 16%). A total of 895 (18%) patients received results from multiple tests within a median time of 76 days. A subset of 2,877 patients tested by one laboratory and with available treatment and survival information was used for biomarker validation and analysis. Twenty-seven combinations of alterations in 12 genes across 13 diseases were associated with prognosis including ARID1A in Brain cancer (HR=0.2, p<10⁻⁶) or NFI in Ovarian cancer (HR=2.06, p<5.10⁻³). The rate of TT matching was high in approved settings (88% of EGFR-mutated lung cancer or KIT-mutated Gastro-Intestinal Stromal Tumors), but low in settings with alternate TT (19% of ERBB2-mutated colorectal cancers). Furthermore, out of 239 combination of genes (N=25), cancer type (N=17) and treatment (N=34) eligible for investigation (N>10 patients in each arm), we identified 20 significant associations (log-rank p-value < 0.01). While most were also prognostic, ARID1A mutations predicted response to oxaliplatin in Gastric cancer (HR=0.17, p<5.10⁻³) and PIK3CA mutations predicted poor response to docetaxel in Breast cancer (HR=7.7, p<3.10⁻³).

ARMOR faithfully captures a snapshot of a molecular oncology clinic using automated and real time collection of both molecular and clinical information. Despite inherent limitations, including phenotypic depth and heterogeneity of the source of the molecular data, ARMOR provides a platform to evaluate the practice of precision oncology at a high-level, rapidly validate matching therapy findings in an independent, real world patient population, document patients’ trajectories, develop predictive models and test novel therapeutic paradigms.
Online health information seeking, health literacy, and human papillomavirus vaccination among transgender and gender diverse people

Anthony Pho, PhD, MPH, ANP-C\textsuperscript{3,8}, Suzanne Bakken, PhD, RN, FAAN, FACMI, FIAHSI\textsuperscript{1,2}, Mitchell R. Lunn, MD, MAS, FACP, FASN\textsuperscript{3,7,8}, Micah E. Lubensky, PhD\textsuperscript{4,8}, Annesa Flentje, PhD\textsuperscript{4,5,8}, Zubin Dastur, MS, MPH\textsuperscript{6,8}, Juno Obedin-Maliver, MD, MPH, MAS, FACOG\textsuperscript{6,7,8}

\textsuperscript{1}School of Nursing, Columbia University, New York, NY, \textsuperscript{2}Department of Biomedical Informatics, Columbia University, New York, NY, \textsuperscript{3}Division of Nephrology, Department of Medicine, Stanford University School of Medicine, Palo Alto, CA, \textsuperscript{4}Department of Community Health Systems, School of Nursing, University of California, San Francisco, San Francisco, CA, \textsuperscript{5}Alliance Health Project, Department of Psychiatry, School of Medicine, University of California, San Francisco, San Francisco, CA, \textsuperscript{6}Division of Gynecology, Department of Obstetrics and Gynecology, Stanford University School of Medicine, Palo Alto, CA, \textsuperscript{7}Department of Epidemiology and Population Health, Stanford University School of Medicine, Palo Alto, CA \textsuperscript{8}The PRIDE Study/PRIDEnet, Stanford University School of Medicine, Palo Alto, CA,

**Introduction** Transgender and gender diverse (TGD) people report a higher prevalence of poor health and experience stigma and discrimination in healthcare, resulting in poor health outcomes from delaying necessary health care. Although health information seeking and eHealth literacy (the ability to use electronic health information to make health decisions) has been explored among cisgender sexual minority people, these topics have not been explored among TGD people (e.g., individuals whose gender identities or gender expressions that may not align with those commonly associated with their sex assigned at birth).[1] There is a paucity of research that explores online health information seeking among TGD people and how this may be associated with health decision-making. The purpose of this study is to describe online health information seeking among a sample of TGD people compared with cisgender sexual minority people to explore associations with HPV vaccination, and whether general health literacy and eHealth literacy moderate this relationship. **Methods** We employed a cross-sectional design to explore the association of online health information seeking and HPV vaccination among TGD and cisgender sexual minority people. Between February and May 2020, we launched an online survey to an existing cohort. Invitations to participate were sent to 17,036 participants in The Population Research in Identity and Disparities for Equality (PRIDE) Study, a longitudinal, U.S.-based, national health study of sexual and gender minority people. The PRIDE Study recruits adults aged 18-years and older, who are English speaking, reside in the U.S. or its territories, and who self-identify as a sexual and/or gender minority person. We employed logistic regression to model the association of health information seeking on the vaccination outcome and interaction terms were entered to test for moderating effects of general and eHealth literacy. Models were adjusted for age, race/ethnicity and education. Alpha for significance was set at .05 with Bonferroni-Holm post hoc correction for adjusted alphas. **Results** The online survey yielded 3,258 responses, of which 1,172 (36%) were classified as TGD and 2,086 (64%) as cisgender. After adjusting for age, race/ethnicity, and education, TGD participants had increased odds of reporting HPV vaccination (aOR=1.5; 95% CI, 1.1-2.2) but decreased odds when they had searched for information about vaccines online (aOR=0.7; 95% CI, 0.5-0.9). TGD participants had over twice the odds of reporting HPV vaccination if they visited a social networking site like Facebook (aOR=2.4; 95% CI, 1.1-5.6). No moderating effects from general or eHealth literacy were observed. **Conclusion** TGD participants had decreased odds of reporting vaccination when they had searched online for vaccine information, but increased odds of vaccination when they reported using social media. Future studies should investigate potential deterrents to HPV vaccination in online health information to enhance its effectiveness and further explore which aspects of social media might increase vaccine uptake among TGD people.

**References**

Leveraging Transformer-based Sequential Sentence Models for Clinical Information Extraction
Ananya Poddar, MS, Venkata Joopudi, MS, Bharath Dandala, PhD, Ching-Huei Tsou, PhD
IBM TJ Watson Research Center, Yorktown Heights, NY

Introduction
Popular machine learning models for clinical information extraction primarily use context within the current sentence without considering the surrounding context in which the sentence appears. As a step toward better document-level understanding, recent work on applying a pre-trained transformer-based language model in a sequential sentence classification (SSC) setting has shown promising results on scientific abstract sentence classification tasks. Inspired by this research, we experiment with two clinical information extraction tasks, namely, note section recognition and clinical summary categorization, that can potentially benefit from inter-sentential context. We demonstrate that a pre-trained transformer-based language model in a SSC setting significantly outperforms that of a single sentence classification setting in these two clinical natural language processing tasks.

Methods
We use expert-labeled annotations from two different datasets for two information extraction tasks: 1) 63,016 sentences from the i2b2 corpora for the note-section recognition task, and 2) 63,594 summary sentences from an internal dataset for clinical summary categorization, to evaluate the usefulness of a SSC setting in comparison to a single sentence classification baseline, widely established in 4.

Results and Conclusion
We achieve state-of-the-art performance on the i2b2 note-section recognition task with a micro-F1 score of 0.822 in the SSC setting, in comparison to 0.710 in the single sentence classification setting. For the clinical summary categorization task, we achieve a micro-F1 score of 0.751 using SSC, an improvement from 0.748 micro-F1 using single sentence classification setting. Our work demonstrates that pretrained transformer-based language models in a SSC setting has the potential to improve state-of-the-art on clinical information extraction tasks. In the future, we plan to explore methods for encoding longer sequences that overcomes context window size limitation and better model interdependencies between consecutive context windows.

References

*We map the labels of the phrase to that of the current sentence and treat this as a sentence classification task.
A Return to Workplace Health Advisor Implementing Public Health Guidance for Safe Re-Opening During a Pandemic

Anita M. Preininger PhD1; Fernando S. Suarez MD1; Jaimie Gorman BS1; Brian Gibbemeyer BBA1; Gretchen P. Jackson MD, PhD1,2

1IBM Watson Health, Cambridge, MA, USA; 2Vanderbilt University Medical Center, Nashville, TN USA

Introduction The COVID-19 pandemic prompted widespread economic shutdowns to contain spread of disease. As science about the novel pathogen emerged, public health agencies provided guidance for returning to work. Return to Workplace Health Advisor (RWHA) was developed to support safe business reopening in accordance with public health recommendations. This poster describes the system implementation and reports early adoption.

Methods RWHA included an employee-facing application that enabled self-screening based on Centers for Disease Control guidance, assessing symptoms, exposures, travel, testing, and risk factors, and generating a work pass (green to allow entrance, red to deny), with capability for employer override (e.g., with medical clearance). The system also featured an employer command center (Fig. 1B), which displayed worksite disease-transmission risk, based on employee-reported diagnoses, symptoms, exposures, and community risk, calculated with data from The Weather Company COVID-19 dashboard, which determined county-level COVID-19 trends using data from national and regional public health authorities. The system was launched in June of 2020 and was incrementally implemented by three medium-large employers and one university in the United States (US). We report initial adoption and work pass results from July to December 2020; data were extracted from the RWHA database. We calculated the proportion of approved passes from the total number of pass requests across all four organizations for the study period, as well as by month. The Western Institutional Review Board determined this study exempt.

Results A total of 75,553 work passes were generated (Jul 2686; Aug 15,895; Sept 15,552; Oct 17,277; Nov 16,123, Dec 8030), with 72,475 (96.2%) of passes allowing workplace access (green) during the study period. The number and proportion of work passes that were not approved for workplace access (red) were: July, N = 78 (2.9%); August, N = 749 (4.7%); September, N = 485 (3.1%); October, N = 723 (4.2%); November, N = 828 (5.1%); December, 218 (2.7%).

Discussion RWHA demonstrated robust adoption and supported thousands of employees in returning to work in the latter half of 2020. The trends in work pass rejection generally paralleled national COVID-19 trends, with the exceptions of an increase in August and decrease in December, likely due university users on an academic schedule.

Conclusion A RWHA with employee self-screening and employer command center allowed thousands of individuals to return to work in accordance with public health guidelines during the global COVID-19 pandemic. Such applications may support businesses maintaining safe workplaces during future infectious disease outbreaks.
Mining for parameters in literature: the case of defining electrical stimulation protocols

Marco Prenassi, PhD¹, Roberto Prandin, Ms¹, Viviana Paolini, Ms¹, Annamaria Caruso, Ms¹, Sara Marceglia, PhD¹
¹Dipartimento di Ingegneria e Architettura, Università degli studi di Trieste, Trieste, Italy

Introduction

Electrical stimulation to the human body is a wide, heterogeneous and ever-evolving research field. Stimulation parameters have wide ranges, due to the type of methodology applied and to the physiological variability between patients. New studies with specific outcomes require to design well-defined protocols; but this task usually involves multiple degrees of freedom due the numerous possible combinations of parameters (electrode location and shape, amplitude, stimulation impulse shape, frequency, pulse width and pulse duration) other than the physiological variability between patients. To help the researchers and the doctors to narrow down the best parameters’ ranges, thus reducing the degrees of freedom of the problem, it is useful to quickly extract and order the data from the latest research. This work aims to help the professionals in this task. We developed a Pubmed-abstract data mining tool that focuses on the four most used electrical parameters: frequency, amplitude, pulse width, and duration, regardless of the specific methodology, and ranks for novelty and subject numerosity the studies that contain that information. The algorithm could be easily customized to accept other parameters (e.g. patient age) and search terms.

Methods

The system architecture is made to be compatible with Python 3.6+. The first module is a custom Pubmed web scraper, based on Requests and BeautifulSoup¹ libraries. This module searches the site using the search terms provided and extract titles, PubMed identifiers (PMID), and abstracts. The second module, based on nltk library² for natural language processing (NLP), tokenizes the text of the abstracts and selects the sentences containing measure units (regular expression string: (\d+.?\d+)s?!(\w+) ). A small custom corpus containing the most used measurement units (i.e. hertz, voltage, ampere, seconds, minutes and relative prefixes) is used to classify the various measurements. Another custom corpus containing the words “patients” and “subjects” and their variations is used to extract the sample size of the study. The algorithm classifies the studies extracted by year, and then by subject number displaying the results as PMID, title and extracted parameters. The algorithm extracts other parameters for further machine learning assisted analysis, as a word frequency counter and, most importantly, orders the extracted sentences as nltk tree with a custom grammar. The human interface is based on PySimpleGui module (version 4.34.0).

Results

To validate the preliminary algorithm, we searched 10 pages of PubMed abstracts each (total: 100 abstracts per query) with the search terms: “Deep Brain Stimulation frequency”, “Transcutaneous electrical nerve frequency” and “Transcranial direct current stimulation voltage”. The algorithm selected the best five articles ordered by year and subject numerosity containing the given parameters. Our team manually ordered the abstracts by year and patient numerosity retaining only the ones with useful parameters (e.g. stimulation frequency, “Hz”) given by the same search terms. The algorithm top five articles matched the manual ranking every time.

Conclusion

Albeit the architecture is in its early prototype stages, it is already usable for preliminary parameter research for targeted electrical stimulation and these functionalities already benefit the design of our protocols. In the future the platform is scalable in two directions: (1) using dedicated modules to search multiple databases and extract even full text articles, where accessible, and implementing a FHIR research-study resource reader. (2) The implementation of machine learning to analyze the natural language sentences surrounding the parameter extraction, to better classify the useful studies.

References

1. Richardson L. Beautiful soup documentation. April. Published online 2007.

Multi-Site Testing of a Prolonged Opioid Prescribing Electronic Clinical Quality Measure Following Elective Primary Total Hip and/or Total Knee Arthroplasties

Avery Pullman, BS1, Mica Curtin-Bowen, BA1, Ania Syrowatka, PhD1,2, Alexandra Businger, MPH1, Michael Sainlaine, MS1, Stuart Lipsitz, ScD1,2, Tien Thai, BS1, Troy Li, BS1, David W. Bates, MD, MSc,1,2 Patricia Dykes, RN, PhD1,2

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

Introduction and Background: The objective of this measure is to provide a tool to assess the prolonged opioid prescribing (POP) rate in opioid-naïve patients following elective primary THA/TKA surgery using routinely collected electronic health record (EHR) data to help drive high quality, evidence-based care. This electronic clinical quality measure (eCQM) will leverage opioid prescription and patient demographic information from EHR systems to compile a risk-adjusted rate of patients who are prescribed opioids for >42 days following elective primary THA and/or TKA procedures. The rate, measured at the clinician group level, is expressed as a percentage where a lower POP percentage is indicative of higher quality care.

Methods and Materials: Alpha and beta testing was conducted at two geographically distant large healthcare systems (referred to as ‘Site 1’ and ‘Site 2’) using EHR data. Alpha testing was comprised of reliability and validity testing to assess the feasibility of implementing the eCQM into EHR systems, and beta testing was conducted to assess the unadjusted and risk-adjusted POP rate across both sites. Site 1 used the EHR vendor ‘Epic’ from six clinician groups from 2016-2019, and Site 2 used the EHR vendor ‘Cerner’ from eight clinician groups from 2017-2019. This measure is stratified, meaning that separate POP rates were calculated for THA and TKA patient populations to reflect the differences in postoperative prescribing practices by procedure type. The same methodology is used for both THA and TKA POP rates.

Discussion: This manuscript outlines the testing of an eCQM intended to assess opioid prescribing practices at two geographically distant sites with distinct patient populations. This study highlights the differences in both data availability and prescribing practices between the two sites. Site 2 recorded notably less data regarding their patient’s opioid episodes (68% in site 2 compared to 100% in site 1) and, in the data recorded, demonstrated patients with a markedly higher rate of prolonged opioid use following surgery.
Introduction:
The use of EHRs has been cited as a key factor in physician burnout, yet few attending physicians, fellows, residents, or medical students ever receive additional didactic or simulated training after their initial onboarding. Current trainees are relatively technology-savvy compared to senior medical providers and could demonstrate new EHR tools. We piloted multispecialty EHR efficiency training targeted for residents, fellows, and faculty attending physicians in graduate medical education (GME) programs.

Methods:
Based on input from faculty and trainees, a 60-minute EHR skills workshop was provided to GME programs at a multi-site healthcare system with a focus on demonstrating skills to improve chart review, documentation, and order entry. Due to social distancing restrictions, workshops were delivered virtually and divided into an instructor demonstration followed by participants practicing the skill in a live EHR environment with support from instructors. Participants received a reference handout and recording of the workshop following the teaching session. Pre- and post-surveys were compared, as well as the number of saved documentation shortcuts in a participant’s account.

Results:
Specialty tailored EHR efficiency workshops were virtually presented to 11 GME program groups (average of 12 participants per session) over fall and winter of 2020, and covered a variety of unique tips (over 20 tips per session) to improve EHR efficiency. Pre-surveys regarding comfort and knowledge of EHR tools were similar between groups. The post-surveys reflected significant increases (p < 0.01) in EHR comfort, efficiency, and tool awareness. The session was recommended by 98% (79 of 81) of the survey participants. Crucially, participants frequently felt this type of training would help reduce burnout (68.5%) and improve work satisfaction (75.4%). In addition, attending providers that participated in training showed a decrease in time spent in chart review (1.8 minutes/day/provider), documentation (3.3 minutes/day/provider), and orders (2.2 minutes/day/provider) post-training. Shortcut use increased in certain subgroups provided training.

Discussion:
Providing EHR Efficiency Workshops is successful in improving providers’ overall comfort and satisfaction with the EHR and appears to show an initial decrease in provider time spent in the EHR. Overall, the program was well received, with both trainees and senior providers requesting the workshops be given earlier and more frequently in training. The forced virtual delivery method limited the ability to correct for technical issues and instructor speed of delivery during the sessions. The reference handout and session recording were appreciated resources by session participants.

Conclusion:
Residents, fellows, and practicing providers find value in learning EHR skills especially related to chart review and documentation efficiency. This type of training should begin early in medical education and reinforced at key intervals throughout training and practice.

References:
Prediction of Five-Year Breast Cancer Recurrence in Women Treated with Neoadjuvant Chemotherapy

Simona Rabinovici-Cohen, MSc1, Xosé M. Fernández, PhD2, Beatriz Grandal Rejo, MD2, Efrat Hexter1, Oliver Hjano Cubelos, PhD2, Juha Pajula, PhD3, Harri Pöllönen, PhD3, Fabien Reyal, Prof. MD2, Michal Rosen-Zvi, PhDL1,4

1IBM Research – Haifa, Mount Carmel, Haifa 3498825, Israel; 2Institut Curie, 26 Rue d’Ulm, 75005 Paris, France; 3VTT Technical Research Centre, Tietotie 4A, Espoo, Finland, 4Faculty of Medicine, The Hebrew University, Israel

Introduction

One of the options for treating locally advanced breast cancer is Neoadjuvant Chemotherapy (NAC), in which chemotherapy and optionally targeted treatment are administered prior to surgery. An important clinical issue to consider when selecting NAC treatment is the likelihood of future cancer recurrence. We present multimodal AI models for predicting cancer recurrence within five years from diagnosis, using both clinical data and multiparametric magnetic resonance imaging (mpMRI) taken prior to treatment. Predicting treatment outcome using multimodal features is an important enabler of precision medicine.

Methods

We work with an anonymized dataset composed of clinical data for 1738 patients and MRI data acquired prior to NAC initiation for a subset of 567 patients. From this dataset, we separated a cohort of 100 patients that had clinical and MRI data for holdout evaluation. The remaining patients were considered for our five-fold cross-validation experiments. Given that we have different sizes of datasets for the different modalities, we divided our model in two branches. The clinical branch was trained using clinical data and the Random Forest classifier. The imaging branch used multiple MRI volumes and a combination of CNNs and classical 3D image processing. We then combined the two branches into one final ensemble model by calibrating the models using Platt’s method and averaging the scores.

The figure shows the imaging mpMRI branch model architecture. It uses multiple MRI volumes of the same study and consists of two components.

Top - Subtraction component in which 7 adjacent MRI slices (3 pre-significant slices, 1 significant slice in which the tumor is the largest, and 3 post-significant slices) form the input to seven 2D-CNNs that have the same weights. The features are aggregated into a 3D-CNN followed by an average global pooling layer.

Bottom - Dixon-ADC component in which three 3D MRI volumes form the input to 3D image processing that generates volumetric features.

Finally, the features from the two components are passed through an FCNN to create the mpMRI score.

Results and Conclusion

We evaluated the individual models, as well as the final ensemble model on cross-validation and on holdout. The mpMRI model achieved 0.700 [0.646, 0.749] AUC and specificity of 0.315 at sensitivity operation point 0.95. In the holdout test, the mpMRI model achieved 0.642 [0.586, 0.695] AUC and 0.213 specificity. The clinical model achieved in cross-validation 0.711 [0.657, 0.759] AUC and 0.263 specificity. On the holdout, the clinical model obtained 0.707 [0.652, 0.756] AUC and 0.196 specificity. The final ensemble model achieved in cross-validation 0.750 [0.698, 0.796] AUC and 0.466 specificity, while in the holdout test it achieved 0.734 [0.680, 0.781] AUC and 0.413 specificity.

In conclusion, we showed that prior to NAC treatment, each individual model shows the ability to predict five-year recurrence, and the multimodal ensemble model offers even better results and clear improvements in specificity.

* Partially supported by the EU’s Horizon 2020 research and innovation program under grant agreement No 780495.
Automatic Data Curation from Unstructured Text

Protiva Rahman, Ph.D., Daniel Fabbri, Ph.D.
Vanderbilt University Medical Center, Nashville, TN

Introduction

Informatics research often relies on data present in unstructured text such as biomedical literature and electronic health record (EHR) notes. As a result, the first step of many projects is curating data from free-text documents. Curation is often a tedious and time-consuming task where curators manually fill out structured data fields, i.e., curation forms. While tools exist for interactive labeling and annotation of medical concepts [1], the underlying models are for niche tasks that cannot be generalized to any given set of fields. In this work, we present our vision for a data curation tool that accelerates and guides curators. Preliminary results of our extraction model show 7% and 16% improvement in accuracy over baseline methods in biomedical literature and electronic notes, respectively.

Methods

Manually curating data into form fields from long documents is a laborious task. To accelerate data curation, we need a tool that learns to automatically extract data and populate curation forms. To this end, we frame the extraction problem as a separate classification task for each form field. The text (from the notes or biomedical article) is the input and the value for the form field is the output or “class”. Keyword search with regular expressions was used as a baseline. The baseline was compared against variations of the BERT algorithm [2], which is a state-of-the-art natural language algorithm. We first use the raw BERT embeddings trained on the Google News corpus. Next, BERT was fine-tuned for one epoch, i.e., one pass through the training dataset, individually for each form field. To overcome the length restriction, the RoBERT [3] model was used. Finally, the performances of BioBERT [4] (trained on PubMed articles) and ClinicalBERT [5] (trained on the MIMIC III notes) were also compared.

Results

Two data curation tasks were used for evaluation. The first is data curation from EHR notes for American Association for Cancer Research GENIE’s BioPharma Collaborative (BPC) [6]. The second use case is data curation for the Preclinical Science Integration and Translation (PRESCIANT) method. The PRESCIANT method aims to extract information from biomedical literature on animal models of human diseases and translate the findings to humans. We report average results across all fields for the two datasets. BERT without finetuning does very poorly. Clinical BERT performs the best for EHR notes, while BioBERT does the best for the literature dataset, as expected.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Keyword Search</th>
<th>BERT (Raw)</th>
<th>BERT</th>
<th>RoBERT</th>
<th>BioBERT</th>
<th>ClinicalBERT</th>
</tr>
</thead>
<tbody>
<tr>
<td>EHR Notes</td>
<td>70.9</td>
<td>21.8</td>
<td>85.3</td>
<td>80.7</td>
<td>86.7</td>
<td>86.8</td>
</tr>
<tr>
<td>Biomedical Literature</td>
<td>53.0</td>
<td>30.8</td>
<td>58.7</td>
<td>57.7</td>
<td>60.1</td>
<td>59.2</td>
</tr>
</tbody>
</table>

Conclusion

Data curation is an integral and time-consuming part of informatics research. Accelerating data curation is vital for widening the bottleneck in the research data pipeline. Our preliminary extraction results show promise, but there are multiple avenues for further improvement. Currently, each form field is extracted individually. However, form fields have dependencies which can be learned if they are trained together, improving overall extraction performance. Moreover, extracting the location along with the value can provide more transparency into the model. Allowing curators to seamlessly and symbiotically work with the system is crucial to the success of informatics research.

Acknowledgment: Research reported in this publication was supported by the AACR GENIE BPC grant.

References

Primary Care Providers’ Needs for Usable Clinical Prediction Rule Presentations

Ivan Rahmatullah, MD MPH1,2, Kari A. Stephens, PhD2,3, Allison M. Cole MD2, Andrea L. Hartzler, PhD3.
1Faculty of Medicine, Universitas Airlangga, Surabaya, Indonesia; 2Family Medicine, University of Washington; 3Department of Biomedical Informatics & Medical Education, University of Washington, Seattle, Washington, United States

Abstract

Clinical prediction rules (CPRs) can improve the accuracy of clinical diagnosis and prognosis and increasingly depend on electronic presentation to clinicians. Yet we lack design studies of CPR presentations based on clinician needs that can improve the usefulness of CPRs. This survey study reports on clinicians’ CPR needs (i.e., preferences, motivations, and barriers) to inform human-centered design of CPR presentations in the primary care context.

Introduction:
Diagnostic and prognostic errors prevent patients from getting timely testing, treatment, and optimal health outcomes. CPRs, such as CENTOR score for the diagnosis of streptococcal pharyngitis and Framingham Score to estimate cardiovascular disease risks could improve the accuracy of clinicians’ diagnoses and prognoses. A review study of 400 CPRs identified primary care as the most targeted study settings. However, CPR presentations that are designed to drive high usability for CPR implementation based on clinician needs remain largely unexplored.

Methods: To characterize clinician needs for CPR presentations, we recruited primary care providers (PCPs) from the WWAMI region Practice and Research Network to participate in an online survey. The survey was executed using REDCap supported by the University Washington. The survey asked about PCP characteristics (i.e., type of provider, years of practice, and CPR experience), preferred CPRs, and motivations and barriers to using CPRs.

For CPR preferences, we presented 11 validated and commonly used CPRs in primary care in the US and asked respondents to rank the top three as most useful. Respondents then ranked their top three motivations out of seven options (i.e., CPRs improve diagnostic/prognostic accuracy, are easy to use, save time for diagnosis/prognosis of a disease, avoid unnecessary costs, improve shared decision-making with patients, are recommended by guidelines, are used by many colleagues). Finally, respondents identified barriers to use of CPRs based on seven options (i.e., CPRs disrupt clinical workflow, don’t improve diagnosis/prognosis accuracy, give results that are difficult to interpret, consume unnecessary time during patient consultation, give results that can be difficult to communicate to patients, interfere with clinician autonomy, increase the risk of being sued). We drew these motivation and barrier options from published reviews and qualitative studies. We summarized survey responses with descriptive statistics.

Results: Of the 25 respondents (34.7% response rate), 80% were MDs (others were PAs, DOs, ARNPs, and Residents), 56% had more than 10 years of primary care experience, and 72% had at least moderate CPR experience. CPRs for diagnosis of depression (i.e., PHQ9/PREDICTNL) and stroke risk (i.e., CHADS2) were preferred as the most useful CPRs, both selected by 84% of respondents. The next most preferred CPRs were CENTOR Score and Framingham Score that were both selected by 44% of respondents. The other CPRs (e.g., CURB score Bacterial pneumonia score, BCRAT, Diabetic foot screen tool, PredictAL, Marburg heart score, and Flu score) received between 0 to 32% selections. For motivation, respondents prioritized ‘easy to use’ (64%), improve diagnostic and prognostic accuracy (56%). For the barriers, the most selected items were disruption to clinical workflow (52%), difficulty communicating CPRs to patients (48%).

Conclusion: Finding identified several targets to address PCP needs for successful implementation of CPRs in practice (i.e., preferred CPRs and prioritized motivations and barriers to use CPRs). Highly ranked CPRs in our sample align well with prior research showing stroke risk (i.e., CHADS2) and depression (i.e., PHQ9) as the most preferred and easy to use CPRs. Findings add prioritized barriers (i.e., workflow disruption and communication difficulties). In addition to supporting prior work, findings add a well-defined list of CPR preference, motivation, barriers from a well-defined target population of PCPs who could benefit from CPR presentations in practice. Findings can inform future design research on CPR presentations that are easy to use, avoid workflow disruption (e.g., by integrating CPR presentations into electronic health records) and address patient communication.

References
Informatics to Power Post-COVID Care: A Framework for Patient Care and Secondary Data Use

Sritha Rajupet MD MPH, Rachel Wong MD MPH MBA, Donna Moller RN-BC, Lisa Maldonado, Tricia Weisse, Tahsin Kurc PhD, Janos Hajagos PhD, Hasit Shah MBA, Mary Saltz MD, Joel Saltz MD PhD, Veena Lingam MD
Stony Brook University Hospital, Stony Brook, NY

Introduction
With over 2 million cases of COVID-19 in New York, approximately 30% of patients have persistent symptoms after infection1-2. Stony Brook Medicine (SBM) launched a Post-COVID multidisciplinary clinic to support patients with Post-Acute Sequelae of SARS-CoV-2 infection (PASC) with primary care and specialty services3-4. Data from PASC forms needed to be integrated into a pre-existing SBM COVID-19 Data Commons, with the goal of identifying a cohort for future researchers to investigate clinical characteristics of PASC, symptom course, and their impact on management and treatment.

Methods
To capture Cerner Millennium™ EHR data in a structured format, we identified Powerforms™ and e-Clipboard™ functionalities that enabled integrated EHR forms to be sent via the patient portal. While building these tools, we leveraged Microsoft Forms™ for interim data collection. With the PASC clinic director, we designed and validated the forms, and developed pipelines to extract, transform, and integrate forms data into the COVID-19 Data Commons. Workflows were developed for forms that required urgent vs. non-urgent review, formalized using swim lane diagrams and approved by risk management and the forms committee for patient safety.

Results
Ten Post-COVID Clinic questionnaires were developed, which could be filled out asynchronously through the patient portal or synchronously by clinical staff using Powerforms™. Both options allowed data to be reconciled by clinicians, integrated into clinical documentation and extracted as discrete data elements. As of July 14, 2021, the SBM Data Commons has incorporated 261 Post-COVID form responses using MS Forms, and 121 using the e-Clipboard. Data quality was evaluated by 1) cross validating visits in the scheduling system with data capture of the associated forms in the Cerner Millennium database to make sure that all responses were recorded and 2) by manual review of 10 sample data extracts to ensure data quality by the clinical informatics team.

Discussion
The PASC clinic was a perfect use-case to develop tools and workflows to collect discrete data on patient-reported outcomes over time. The front-end effort of building integrated EHR forms allowed easy integration into the local Data Commons, future generalizability of the EHR forms functionalities and the ability to scale up data-sharing to national databases with minimal pre-processing.

References
Factors Historically Influencing Telehealth Use in Six Medical Specialties:
A Systematic Review and Narrative Synthesis

Pavani Rangachari, PhD1, Swapandeep S. Mushiana, M.S.2, Krista Herbert, M.A.3
1Augusta University, Augusta, GA, USA; 2 University of California, San Francisco, CA, USA; 3 Rowan University, Glassboro, NJ, USA

Background: Recent studies preceding the COVID-19 pandemic in the US, have identified wide variations in telehealth use across medical specialties. This is a problem-of-interest, because the US has historically lacked a standardized set of telehealth coverage & reimbursement policies, which in turn has posed a barrier to telehealth use across all specialties. Although all medical specialties in the US have been affected by these macro (policy-level) barriers, some specialties have been able to integrate telehealth use into mainstream practice, while others are just gaining momentum with telehealth during COVID-19. Despite accelerated use of telehealth services during the pandemic, owing to the temporary removal of federal coverage restrictions, uncertainties remain regarding future sustainability. Since macro (policy-level) factors do not serve to explain the variation in telehealth use across specialties, it would be important to examine meso (organizational-level) and micro (individual-level) factors historically influencing telehealth use across specialties, to understand underlying reasons for the variation and identify implications for widespread sustainability.

Methods: This paper draws upon the existing literature to develop a conceptual framework on macro-meso-micro factors influencing telehealth use within a medical specialty. The framework is then used to guide a systematic review and narrative synthesis of the specialty-level telehealth literature, to identify factors historically influencing telehealth use across six specialties in the US, including three “lower-using” specialties (Allergy-Immunology, Gastroenterology, Family Medicine) and three “higher-using” specialties (Cardiology, Psychiatry, Radiology). Article searches were conducted on PubMed. The PRISMA checklist was used to guide reporting of literature reviewed. Three reviewers worked to develop a preliminary synthesis, identify eligibility criteria, explore themes in the data, and assess robustness of final synthesis.

Results: Fifty-three articles were reviewed across six medical specialties. The review identified 12 factors across 3 layers, including: 1) macro-layer (policy-level, legal, structural factors), 2) meso-layer (specialty-level historical telehealth rationale, hospital-organizational, specialty-society, treatment, technological, research, cultural factors) and 3) micro-layer (individual-level provider-and-patient-specific factors). There was limited variation across the six specialties, with respect to factors influencing telehealth use in the macro layer. However, distinct themes emerged among “lower-using” and “higher-using” specialties in meso and micro layers. For example, the historical rationale for telehealth use in “lower-using” specialties, was ‘improving access to care.’ By contrast, “higher-using” specialties had progressed beyond using telehealth to improve access, to using it to ‘improve patient experience/outcomes,’ ‘reduce costs,’ or ‘promote population health’ (Triple Aim). Another key finding was that among “higher-using” specialties, the specialty-societies and hospital-organizations in the meso-layer, proactively promoted telehealth use by influencing both macro-layer factors (e.g., advocating for consistent payment policies) and micro-layer factors (e.g., enabling provider practices to be more tech-savvy & patient-centric).

Conclusions: By identifying a comprehensive set of policy, organizational, individual (and interaction) factors influencing telehealth use across six medical specialties, this review addresses a gap and provides a foundation for future research. Importantly, it identifies: 1) strategies for reducing variation in telehealth use across medical specialties, and 2) implications for ensuring widespread sustainability in the post-pandemic era. For example, to secure sustainable support for telehealth services from hospitals and payers, providers in “lower-using” specialties could undertake telehealth initiatives in line with the Triple Aim framework (e.g., allergy providers could develop capacity for remote monitoring of asthma management to demonstrate reduction in hospitalizations and costs). Likewise, specialty societies and hospital organizations could play a significant role in promoting telehealth sustainability, by advocating for better telehealth reimbursement from public and private payers; and educating & training providers in the effective design and implementation of telehealth services.
Generalizable Gated Recurrent Neural Network based model to predict COVID-19 patient outcomes on admission

Laila Rasmy, MSc¹, Bijun Sai Kannadath, MBBS, M.S², Masayuki Nigo, M.D³, Ziqian Xie, Ph.D¹, Bingyu Mao, MA¹, Khush Patel, MBBS¹, Yujia Zhou, MSc¹, Xiao Dong, MSc¹, Hua Xu, PhD¹, Degui Zhi, PhD¹

¹School of Biomedical Informatics, UTHealth, Houston, TX; ²College of Medicine, University of Arizona - Phoenix, AZ; ³McGovern Medical School, UTHealth, Houston, TX.

Introduction

Though multiple machine learning based predictive algorithms have been developed using secondary electronic health record (EHR) data, these models have rarely been widely implemented or validated in practice despite their high reported accuracy. During the pandemic and by July 2020, Wynants et al reviewed 107 COVID-19 prognostic predictive models, out of which they only found one promising model that deserves further validation. The common challenge for the majority of the studies is the high risk of bias. Training and evaluating models on large diverse data, reporting calibration results, and external validation are among the methods that can help reduce the risk of model bias. Additionally, such models mostly need complicated data preprocessing pipelines that impact their transferability, reliability, and sustainability. In our study, we propose a gated recurrent neural network (GRU) based algorithm that utilizes EHR structured data to predict patient outcomes with minimal need for data preprocessing. We trained our model using a cohort of 243,785 patients from 85 different health systems available through the cloud-based, de-identified COVID-19 patient data derived from the Cerner® Real-World Dataset (CRWD). For external validation and transferability evaluation, we evaluated the model on two held-out hospitals' data from Cerner RWD as well as a cohort of 36,140 patients extracted from de-identified patient data derived from the Optum® COVID-19 EHR-based dataset (2007-2020) was employed for external validation of our model.

Methods

We extracted all patient information available on and before the day of admission, including demographics, diagnosis, medication, procedures, laboratory results, and observations. We trained our model embedding layer on standards commonly available in EHRs such as ICD9/10CM, ICD9/10PCS, SNOMED-CT, LOINC, Multum, CPT-4, and HCPCS. We split our primary cohort into training, validation, and test sets using the ratio of 7:1:2 respectively. We reported the discriminatory prediction accuracy measured by AUC on a held-out test set of 48,781 from 85 health systems. For external validation, we evaluated the model on two held-out hospitals from different Census regions, Hospital 1 from the south and Hospital 2 from the west. We further tested the transferability of our proposed model on the OPTUM cohort.

Results & Conclusion

Our GRU-based model achieved a high discriminatory prediction accuracy measured by AUC as well as good calibration for different tasks including in-hospital mortality (93%), need for mechanical ventilation (93%), long hospital stays (>7 days) (87%), as well as a good calibration. External validation showed a consistently good performance with AUC of 91% and 97% for Hospital 1 (3,469 patients) and Hospital 2 (706 patients) in CRWD, respectively, and 90% on OPTUM cohort for in-hospital mortality. Our results were consistently better than baseline models such as logistic regression by around 3%. To the best of our knowledge, We are the first to train an RNN-based model utilizing medical code embedding for the prediction of COVID-19 outcomes from EHR data on admission. Trained on a large heterogeneous dataset from 85 health systems, our model showed one of the highest reported prediction accuracy, transferability, and consistently good performance on multiple external datasets.

Acknowledgments

L.R. and X.D. are supported by the UTHealth Innovation for Cancer Prevention Research Training Program Pre-Doctoral Fellowship (CPRIT Grant No. RP160015)

References

Cardiac patients’ and healthcare providers’ telehealth experiences during the COVID-19 pandemic in Queens, New York

Meghan Reading Turchioe, PhD, MPH, RN\textsuperscript{1}, Alexander Volodarskiy, MD\textsuperscript{2}, Jessica S Ancker, PhD\textsuperscript{3}, Joshua Vapnik, MD\textsuperscript{2}, Sushant Sunkaraneni, MD\textsuperscript{3}, David Slotwiner, MD\textsuperscript{1,2}

\textsuperscript{1}Weill Cornell Medicine, Dept. of Population Health Sciences; \textsuperscript{2}NewYork Presbyterian Hospital-Queens; \textsuperscript{3}Vanderbilt University Medical Center, Dept. of Biomedical Informatics

Introduction: Emerging research suggests that federal policy changes aiming to expand access to telehealth in response to the COVID-19 pandemic may be widening health disparities among racial/ethnic minority and low income individuals\textsuperscript{1}. Cardiac patients are more likely to contract and suffer worse outcomes from COVID-19, and racial/ethnic minority and low-income cardiac patients are at the highest risk\textsuperscript{2}. An understanding of cardiac patients’ and providers’ experiences with telehealth is lacking, but critically needed to tailor telehealth delivery in ways that close disparities.

Objective: To compare experiences with telehealth during the COVID-19 pandemic among cardiac patients and providers at a hospital in Queens, New York, a highly diverse county that has been severely impacted by COVID-19\textsuperscript{3}.

Methods: We conducted a cross-sectional study from August-October 2020 involving a one-time survey with questions about experience with two common types of telehealth, video visits and remote patient monitoring, based on the Technology Acceptance Model. Data from remote monitoring included blood pressure, blood glucose, weight, and cardiac rhythms from wearables or implantable cardiac devices. Cardiologists and internal medicine physicians were recruited by email and completed the survey online. A convenience sample of cardiac patients who recently received care were recruited and completed the survey by phone to avoid biasing the sample towards higher technology comfort. An interpreter service was used as needed. We compared cardiac patients and healthcare providers responses using chi-squared or Fisher’s exact tests as appropriate. The Weill Cornell Medicine IRB approved this study.

Results: Fifty-four participants completed the study. Most providers (n=26) were cardiologists (72%). Patients (n=28) had a mean age of 61.6 (SD 17.7) years, and were 64% female, 25% Black, 39% Asian, 21% Hispanic/Latino, and 28% non-English speaking. Additionally, 32% of patients reported inadequate financial resources and 36% received Medicaid insurance. More than half of participants had experience with both video visits and remote monitoring (Table 1). Compared to patients, significantly more providers had experience with video visits before March 2020. Providers reported significantly higher ease of use and comfort making medical decisions over video visits, while patients reported significantly higher ease of use and perceived treatment quality using remote monitoring data.

| Table 1. Comparing telehealth experiences during COVID-19 among cardiac patients (n=28) and providers (n=26) |
|-------------------------------------------------|----------------|----------------|----------------|----------------|----------------|----------------|
| Variable                                        | Video visits | Remote monitoring | p value | Video visits | Remote monitoring | p value |
| Any experience                                  | 18 (64%)     | 22 (85%)         | 0.12   | 16 (57%)     | 22 (85%)         | 0.04   |
| Experience before March 2020                   | 0 (0%)       | 12 (46%)         | <0.001 | 11 (39%)     | 8 (31%)          | 1.00   |
| COVID-19 increased acceptance                  | 14 (50%)     | 13 (50%)         | 0.89   | 11 (39%)     | 17 (65%)         | 0.02   |
| High perceived ease of learning to use*        | 7 (39%)      | 16 (73%)         | 0.07   | 4 (25%)      | 12 (55%)         | 0.02   |
| High perceived ease of use overall*            | 8 (44%)      | 18 (82%)         | 0.02   | 11 (69%)     | 4 (18%)          | <0.001 |
| High perceived better treatment*               | 8 (44%)      | 14 (64%)         | 0.37   | 11 (69%)     | 6 (27%)          | 0.01   |
| High perceived usefulness overall*             | 8 (47%)      | 16 (73%)         | 0.19   | 12 (75%)     | 15 (68%)         | 0.33   |
| High comfort making major treatment decisions* | 4 (22%)      | 13 (59%)         | 0.04   | 12 (75%)     | 15 (68%)         | 0.33   |

*Among participants with any experience

Conclusion: Patients perceive remote monitoring significantly improves the care they receive and are comfortable working with their providers to make major treatment decisions based on the data obtained from remote monitoring. Patients are less confident than providers that video visits are useful for their care and they are less comfortable making major treatment decisions during a video visit. Improved technical support and patient education regarding what types of medical decisions may be appropriate to reach during a video visit may improve patient acceptance.

References
Piloting FDA SHIELD’s LOINC to In Vitro Diagnostic (LIVD) Specification in Five Medical Centers: Implications for Interoperability

Greg Rehwoldt, PhD, MBA¹, Raja Cholan, MS², Andrew Sills, BS³, I. Khalil Appleton, BS⁴, Tim Williams, BS⁴, Natalie Scott, BS⁴, Gregory Pappas, MD, PhD⁵

¹Deloitte Consulting LLP, ²Food and Drug Administration

Introduction: To address laboratory data challenges, the Food and Drug Administration (FDA) sponsored the Systemic Harmonization and Interoperability Enhancement for Laboratory Data (SHIELD) workgroup focused on improving the interoperability of IVD data. One of SHIELD’s solutions is the LIVD (LOINC to IVD) mapping specification, which harmonizes how IVD is represented using LOINC and SNOMED CT. The FDA is funding a SHIELD demonstration program to assess the effectiveness of the LIVD specification in five medical centers’ laboratory settings. In this paper our goals are to (1) describe the laboratory coding information collected from the five pilot sites, (2) assess the extent to which LOINC codes in the LIVD catalog files from IVD manufacturers compare to LOINC codes chosen in information systems in five pilot medical centers, and (3) discuss lessons learned and implications for the SHIELD workgroup.

Methods: We sought to identify gaps and similarities between LOINC codes from the manufacturer LIVD files versus LOINC codes from the medical centers. First, we recruited five medical centers to participate in FDA’s SHIELD pilot. Each medical center was asked to extract 100 LOINC Codes from LIS systems for prioritized tests of interest focused on high risk conditions and SARS-CoV-2 / COVID-19. For each selected test (e.g., SARS-CoV-2 RNA COVID-19), we collected the following data elements: test names/descriptions (e.g., SARS coronavirus 2 RNA [Presence] in Respiratory specimen by NAA with probe detection), associated instruments (e.g., IVD Vendor Model), and LOINC codes (e.g., 94500-6). Next, we collected manufacturer LIVD catalogs containing the LOINC codes per IVD instrument per test from manufacturers. Finally, we compared the overlap between the LOINC codes from manufacturers versus medical centers for the same tests. We evaluated the gaps and similarities by the two sources (IVD manufacturers versus medical centers) to determine what tests/codes were mismatched.

Results: Five academic medical centers were onboarded from five distinct states throughout the United States, ranging in size from 500 to 1,500 hospital beds. Three IVD Manufacturers have submitted 15 LIVD catalogs representing 26 distinct devices, 6,956 tests, and 686 LOINC codes. Table 1 summarizes the data collected from the pilot sites and the matches/mismatches between LOINC codes from the manufacturer LIVD files versus LOINC codes from the medical centers.

Table 1. Medical Center Pilot Site Laboratory Information System (LIS) Data Reporting and Interoperability Matches and Mismatches Between / Within Manufacturers & Medical Centers

<table>
<thead>
<tr>
<th>Medical Center Pilot Site</th>
<th># of Tests / LOINC Codes Collected</th>
<th># of Manufacturer to Pilot Site Matches between LOINC Codes</th>
<th># of Manufacturer to Pilot Site Mismatches between LOINC Codes</th>
<th># of Distinct IVDs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Abbott Biomerieux Roche</td>
<td>Abbott Biomerieux Roche</td>
<td></td>
</tr>
<tr>
<td>Hopkins</td>
<td>510</td>
<td>N/A N/A 91</td>
<td>N/A N/A 106</td>
<td>17</td>
</tr>
<tr>
<td>Intermountain</td>
<td>70</td>
<td>1 N/A</td>
<td>7 12 N/A</td>
<td>19</td>
</tr>
<tr>
<td>Miami</td>
<td>74</td>
<td>N/A N/A 50</td>
<td>N/A N/A 4</td>
<td>17</td>
</tr>
<tr>
<td>UNMC</td>
<td>526</td>
<td>N/A 29 N/A</td>
<td>N/A 0 N/A</td>
<td>32</td>
</tr>
<tr>
<td>Yale</td>
<td>139</td>
<td>N/A N/A 124</td>
<td>N/A N/A 17</td>
<td>19</td>
</tr>
<tr>
<td>Total</td>
<td>1319</td>
<td>7 30 265</td>
<td>7 12 127</td>
<td>104</td>
</tr>
</tbody>
</table>

Conclusion: The five medical centers vary in how they organize, categorize, and store LIS catalog information. We identified mismatches in how medical centers use LOINC to encode laboratory tests compared to how IVD Manufacturers use LOINC to encode the same laboratory tests in the LIVD catalogs. A summary of key findings include the following: (1) Medical Center LIS test catalogs include data quality inaccuracies with LOINC, such as using codes that are not proper LOINC codes, deprecated LOINC codes, discouraged LOINC codes, and trial LOINC codes; (2) Medical Center LIS catalogs can contain duplicative tests or tests that change meaning over time (identifiers/descriptions change over time); (3) Medical centers expressed that the LIVD catalog is helpful as a centralized platform, taking away LOINC coding guesswork and reducing LOINC coding variation between systems. There is potential for the LIVD catalogs to help improve semantic interoperability and data quality of LIS data; there is also room for improvement of the LIVD catalog data elements and accessibility of LIVD files by labs. Enhanced program support for SHIELD is necessary to effectively rollout LIVD in medical centers and promote lab data interoperability nationally. Suggested next steps include additional collaboration among key players in the laboratory ecosystem including standards development organizations, device manufacturers, and medical centers.
Usability Study of a Decision Support Website to Support People Living with Cystic Fibrosis in Shared Decision Making about Lung Transplant

Nick Reid, MHI1, Kathleen J. Ramos, MD MSc2, Mara R. Hobler, PhD2, Lauren E. Bartlett, BS2, Siddhartha G. Kapnadak, MD2, Andrea L. Hartzler, PhD1
1Department of Biomedical Informatics and Medical Education, School of Medicine, University of Washington, Seattle, WA, USA; 2Division of Pulmonary, Critical Care, and Sleep Medicine, Department of Medicine, University of Washington, Seattle, WA, USA

Introduction: Although individuals with advanced cystic fibrosis (CF) lung disease can benefit significantly from lung transplant, many are apprehensive, lack information, and experience barriers to transplant discussions with providers. Patients with advanced CF lung disease often experience ambivalent and complex feelings regarding lung transplant, with some wanting all available transplant information and others preferring to moderate the information they receive. There is a need for shared decision-making tools that help this range of patients prepare for discussions with providers about lung transplant.

Our preliminary focus groups in people with CF indicated that a shared decision-making tool should include didactic educational materials, patient stories, and frequently asked questions (FAQ). Despite the potential benefits of such content, it is unknown what design options CF patients find most usable. Some shared decision-making tools adopt an “author-driven” design that guides readers through content and recommends content that is appropriate. Conversely, “reader-driven” designs present content uniformly and allow the reader to choose independently. This study aimed to delineate which design people with CF find most usable and useful for learning about lung transplant.

Methods: We conducted individual sessions with CF patients to compare the usability of two alternative website prototypes. The reader-driven prototype displayed each content type (didactic, stories, FAQ) as separate sections, allowing participants to freely browse the available content. The author-driven prototype guided the user to complete a survey before displaying recommended content from all three types. During each session, participants interacted with both author-driven and reader-driven prototypes to complete scenario-based usability tasks, then completed the System Usability Scale (SUS) for each prototype. We summarized SUS scores with descriptive statistics, and then compared scores between prototypes with paired t-tests. Website usage logs for each prototype were compared to identify which content type was viewed first by participants and frequencies were summarized.

Results: Fourteen people with CF who had not received a lung transplant completed the study. Table 1 compares SUS scores and frequency counts from website usage logs. The mean SUS score for the reader-driven prototype was significantly higher than the author-driven (90.0 vs. 81.6, p=0.002.) Analysis of website usage logs showed that participants first viewed a wider variety of content types in the reader-driven prototype when compared to the author-driven prototype.

Table 1. Comparison of two prototype shared decision-making tools, author-driven and reader-driven

<table>
<thead>
<tr>
<th>System Usability Scale Score Mean (SD)</th>
<th>Content type viewed first (n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>Usability</td>
</tr>
<tr>
<td>Author-Driven</td>
<td>81.6 (9.9)</td>
</tr>
<tr>
<td>Reader-Driven</td>
<td>90.0 (8.7)</td>
</tr>
</tbody>
</table>

Conclusion: This usability study demonstrated that people with CF, when learning about lung transplant, found a reader-driven design more usable. This design allowed them increased control to access more varied content within a shared decision-making tool. The reader-driven prototype had a significantly higher mean SUS score than the author-driven prototype (90.0 vs 81.6, p=0.002), indicating that participants perceived the reader-driven prototype more usable. These results can inform the design of a shared decision-making tool for people with CF considering lung transplant.

References
Relationship Between Social Determinants of Health and Total Joint Arthroplasty Complications

Lauryn Remmers¹, Edward Kalpas, MD, MAS, MPH¹,², Priya Radhakrishnan, MD, FACP², Anita Murcko, MD, FACP, FAMIA¹

¹Arizona State University, Tempe, AZ; ²HonorHealth, Scottsdale, AZ

Research Objective
Social determinants of health (SDOH) are the conditions in one’s living environment that affect health, functioning, and quality of life¹. Total joint arthroplasty (TJA) is a surgical procedure to replace a damaged joint with an artificial joint. TJA complications include acute myocardial infarction, pneumonia, sepsis, surgical site bleeding, pulmonary embolism, or periprosthetic joint infection². Previous research demonstrates that Black race, Hispanic ethnicity and poverty were negatively associated with TJA outcomes in veterans³. The goal of this mixed methods quality improvement study is to determine if SDOHs affect TJA complications at a health system in the Phoenix metropolitan area.

Methodology
For this study, records from patients who underwent hip or knee TJAs at any of the five system facilities between 2/2019-2/2020 were included. Demographics and clinical data were extracted from the electronic health record (EHR) via Midas® Care Management with SDOH variables from case manager notes corresponding to food, utilities, housing and transportation insecurities, and interpersonal safety. Complications were identified using ICD-10 codes. SDOH for individuals with and without complications were compared. A multinomial logistic regression was performed in SPSS to identify significant variables. Semi-structured interviews with case managers (n=2), orthopedic surgeons(n=5), and primary care physicians (n=4) were performed to explore care team interactions with SDOH. Interview notes were coded and analyzed based on response frequency and themes.

Results
Of 2,520 patients who underwent TJA, 50 (1.98%) experienced a TJA complication. Of those, 38% screened positive for an SDOH. For those without a TJA complication, 27% screened positive for an SDOH (p=0.093). Most interview participants identified a correlation between socioeconomic status and surgical outcomes. They also recognized that language barriers for Spanish-speaking individuals and family involvement post-discharge are significant factors in TJA outcomes.

Table 1. Multinomial logistic regression model of predictors of surgical complications after total joint arthroplasty

<table>
<thead>
<tr>
<th>Covariate</th>
<th>B</th>
<th>Exp(B)</th>
<th>Std. Error</th>
<th>p-value* &lt;0.05 significant</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>0.008</td>
<td>1.008</td>
<td>0.015</td>
<td>0.596</td>
<td>0.979-1.038</td>
</tr>
<tr>
<td>No SDOH Factors</td>
<td>-0.339</td>
<td>0.713</td>
<td>0.313</td>
<td>0.279</td>
<td>0.386-1.315</td>
</tr>
<tr>
<td>Facility A</td>
<td>0.120</td>
<td>1.128</td>
<td>0.577</td>
<td>0.835</td>
<td>0.364-3.497</td>
</tr>
<tr>
<td>Facility B</td>
<td>0.973</td>
<td>2.645</td>
<td>0.490</td>
<td>0.047*</td>
<td>1.013-6.910</td>
</tr>
<tr>
<td>Facility C</td>
<td>0.851</td>
<td>2.342</td>
<td>0.470</td>
<td>0.070</td>
<td>0.932-5.885</td>
</tr>
<tr>
<td>Facility D</td>
<td>0.200</td>
<td>1.222</td>
<td>0.466</td>
<td>0.667</td>
<td>0.490-3.048</td>
</tr>
<tr>
<td>Not Caucasian</td>
<td>0.049</td>
<td>1.050</td>
<td>0.611</td>
<td>0.936</td>
<td>0.317-3.479</td>
</tr>
</tbody>
</table>

Conclusions
This single system mixed methods retrospective quality improvement study demonstrates that patients who screen positive for an SDOH are more likely to experience a TJA complication. We recommend that SDOH assessments be obtained for all patients undergoing TJA, be available to care teams, and be incorporated into care plans to improve outcomes.

References
An Evaluation of Solor and ANF Assignment to HL7 Knowledge Artifacts Using High Definition Natural Language Processing

Melissa P. Resnick, PhD, MLS, MS¹, Frank LeHouillier, MA¹, Steven H. Brown, MD², Keith Campbell, MD, PhD², Diane Montella, MD², Peter L. Elkin, MD¹²

¹University at Buffalo, Buffalo, NY, USA,
²U.S. Department of Veterans Affairs

Introduction

Standards, such as terminologies, assist with important tasks including, natural language processing (NLP) and semantic interoperability. Unstructured notes written by healthcare providers in the electronic health record (EHR) are often difficult to access. This unstructured text is critical for providing healthcare, data analysis, and research. One can access this text by utilizing a terminology together with an NLP program and a standard statement model, such as, ANF¹. Solor is an integrated terminology constructed in collaboration with the U.S. Veterans Affairs (VA)². Solor combines SNOMED CT (for diseases, findings, and procedures), Logical Observation Identifiers, Names, and Codes (LOINC) (for laboratory test results), and RxNorm (for medications)²,³. A single integrated terminology, such as Solor, allows clinical data to flow between clinical documentation, decision support applications, and physician order entry at the point of care. ANF is a type of information model designed to be independent of the content in the clinical statements. The objective of ANF is to provide a simple and consistent information model for clinical statements. This makes it easier for analysts to understand the data and the manner in which it is stored. The HD-NLP program, previously known as HTP-NLP, is software developed at the University at Buffalo⁴. The HD-NLP is a name entity recognition type of program⁵. The unstructured text from the EHR is input into the HD-NLP program, which then uses a terminology to output tagged text⁶. The HD-NLP can also provide Solor terms in the ANF output.

Methods

HD-NLP was used to process 694 clinical narratives previously modeled by human experts into Solor and ANF. The HD-NLP output was compared to the expert gold standard for 20% of the sample. Each clinical statement was judged "correct" if HD-NLP output matched ANF structure and Solor concepts, or "incorrect" if any ANF structure or Solor concepts were missing or incorrect. Judgements were summed to give totals for "correct" and "incorrect".

Results

113 (80.7%) correct, 26 (18.6%) incorrect, and 1 error. Inter-rater reliability was 97.5% with Cohen’s kappa of 0.948. of 80% with these complex constructions needed for input to clinical decision support.

Conclusion

This iteration of the HD-NLP program provided an acceptable level of accuracy.

References


Are Smartphone-based Interconnected Personal Health Records Achieving Their Promise?

Tera L. Reynolds, MPH, MA¹, Meghna Kaligotla¹, Kai Zheng, PhD¹
¹University of California-Irvine, Irvine, CA, USA

Introduction

Smartphone-based interconnected personal health records (PHRs) are mobile applications (apps) “through which patients can maintain and manage their health information (and that of others for whom they are authorized) [from multiple electronic health record systems] in a private, secure, and confidential environment.” This type of PHR has the potential to integrate a patient’s clinical data such as laboratory test results from multiple healthcare organizations and to make that data accessible to other health apps to potentially improve their user experience (e.g., by automatically importing medication information). Ultimately, this empowers patients through opportunities to leverage their clinical data (e.g., to improve their health). Through this study, we aimed to investigate the landscape of third-party apps that claim to be able to connect to Apple Health Records, the most commonly available smartphone-based interconnected PHR, or to CommonHealth, the Android-based app developed to serve a similar purpose as Apple Health Records.

Methods

We first generated a list of health-related iOS apps using the Apple App Store’s ranking of the top apps in their Health and Fitness and Medical categories as of October 2020. This resulted in a list of 601 unique apps. In addition, we did targeted searches using keywords based on the initial Apple-defined use cases of apps that may benefit from using clinical data in their functioning—medication, disease management, nutrition, and research—resulting in 199 additional unique, health-related apps. For the final list of 800 apps, we extracted pre-defined elements, including app name and price. To identify relevant apps—those that can use Apple Health Records’ API to import clinical data—two researchers reviewed the title and short description of each app to determine the ‘potentially relevant’ apps (i.e., those that may use clinical data). We identified 81 potentially relevant apps and extracted full descriptions for these apps. We then performed a keyword search of the descriptions using terms such as “apple health” and “apple health records,” and identified 43 apps with at least one keyword. We followed a similar process for the Google Play Store in January 2021, resulting in a list of 611 health apps; three of the descriptions had the keyword “commonhealth.” Finally, TLR and MK manually reviewed the full descriptions of potentially relevant apps. In all phases of manual review, any disagreements were resolved through discussion.

Results

Thirteen of the 800 iPhone health and medical apps reviewed were not available in the U.S., had no description available, and were, thus, excluded. Among the remaining 787 apps, there were only five apps that claim to be able to import clinical data from Apple Health Records (0.6%). These apps fell into four different categories: medication management (MediSafe), condition self-management (Livongo), standalone telehealth (MDLive), and research (All of Us Research and All of Us Research Program). In addition, while three health apps from the Google Play Store were identified as potentially relevant, upon manual review of the descriptions, none of these apps appear to be able to connect to the CommonHealth PHR app.

Conclusion

While there is great potential in computable electronic medical records via apps connected to smartphone-based interconnected PHRs, among the most popular health apps in Apple’s app store, few seem to offer this feature. This suggests that the vision for a robust digital health ecosystem that offers patients flexible solutions to better meet their health and healthcare needs has yet to come to fruition. In addition, care needs to be taken to avoid digital health disparities based on the type of smartphone one owns.

References

An Interactive Data Extraction System to Create the Living Systematic Reviews and Meta-Analysis

Irbaz Bin Riaz, MD, MS  1,4*, Huan He, PhD  2*, Syed Arsalan Ahmed Naqvi, MBBS  3  
Rabbia Siddiqi, MBBS  3, Noureen Asghar MBBS  3,  
M. Hassan Murad, MD, MPH  4, Hongfang Liu, PhD  2  
1Department of Oncology, Mayo Clinic, Phoenix, AZ, USA  
2Department of AI and Informatics Research, Mayo Clinic, Phoenix, AZ, USA  
3 Dow University of Heath Sciences, Karachi, Pakistan  
4 Mayo Clinic Evidence Based Practice Center, Mayo Clinic, Rochester, MN, USA

Introduction: It takes months to years for conducting rigorous systematic review (SR) and meta-analysis (MA), and it involves developing a search strategy, screening for relevant citations, extracting and analyzing data and ironically the review can be outdated as soon as it published. The current process of creating SRMAs is inadequate to keep pace with rapid influx of evidence as seen in COVID-19 pandemic and dynamic fields like Oncology. Thus, living SRs (LSRs)—which are updated as soon as new evidence becomes available—are necessary to overcome the limitations of conventional reviews. Previously, we have described semi-automating the screening process and analyses, and here we describe our innovative method for data extraction to decrease the effort for creating and maintain LSRs.

Methods: We designed three major modules to facilitate the data extraction: the outline module (Fig. 1(A)), the table module (Fig. 1(B)), and the interactive extractor (Fig. 1(C)). The purpose of outline module is to create a skeleton of tables for data extraction. In the outline module, the users specify the outcomes of interest as well as the structure of the data format required to analyze each outcome. Users are provided with a range of standard options to create summary and data tables and are also provided with the flexibility to generate new variables as necessary. The tables in the table module are automatically populated with studies to be included in the review with the meta-data such as year of publication, author and journal name and rest of tables are completed with the help of extractor module. The studies are listed row by row. Users could select the “check” option and select each particular study for extraction. When a study row in this table is clicked, the interactive extractor would show its text details, including the abstract and full-text PDF files. The extractor module allows user to extract data from associated text. The user simply needs to highlight the text, right-click on the highlighted text, select which attribute the highlighted text belongs to. Then the highlighted text would be saved as the value of the selected attribute for this study. Thus, the tables generated from these modules can be presented to summarize the results of the systematic review or perform further analyses such as pairwise or network meta-analyses, create summary of findings tables and evidence maps.

![Figure 1](image)

Figure 1. The screenshots of our extraction tool. (A) the outline module; (B) the table module; (C) the interactive extractor.

Discussion: Extracting data for LSRs is particularly challenging as it requires management of the key information from meta data and unstructured free texts such as PDF files and web pages. Although existing tools, such Rayyan and Covidence, work well on independent tasks (e.g., screening, data retravel) and generate data files, linking those isolated data files correctly is challenging and requires tedious manual operations in multiple tools. Thus, to ease the burden of using we created a user-friendly interface which brings together data from multiple sources and facilitates data extraction and creation of summary tables and data tables for analysis. Next steps include integrating this data extraction system with our previously created modules for screening and analyzing the data thereby creating a pipeline for true, living systematic reviews which will be updated in “almost” real time.

* Those authors have contributed equally as co-first authors
AI Assisted Mobile Triage App: Assessment of Time to Decide, Choice of Course of Action, Confidence Level, and Perceived Usability

Adam Rich, RN\textsuperscript{1}, Camille Whicker \textsuperscript{1}, Phung Matthews, PharmD\textsuperscript{1}, Charlene Weir, RN, PhD \textsuperscript{1}; Damian Borbolla, MD, MS \textsuperscript{1}.

\textsuperscript{1}Department of Biomedical Informatics, University of Utah, Salt Lake City, UT

Introduction
With advancements in technology, and the ease with which information is readily available online, individuals are utilizing websites and applications to self-diagnose, prior to seeking medical care/advice in a clinical setting. Artificial intelligence (AI) assisted applications are surfacing, with aims of providing accurate diagnoses and courses of action to their users. In this study, we assess the triage decision and accuracy, time to arrive at a decision, the user’s confidence level in their decision, and the usability of the AI assisted symptom triage application.

Methods
The study followed a mixed method design, assessing participant’s decisions regarding medical course of action after reading four clinical case scenarios. Participants were recruited through a snowball technique. Zoom evaluations were utilized to engage with participants. Each participant read through each scenario before deciding their “course of action”. The outcome measures included: 1) chosen course of action (\textit{Go to the ER/Urgent Care}, \textit{Call your Provider}, or \textit{Stay at home and monitor}); 2) time to arrive at that decision; 3) participant’s confidence level in their decision; and 4) perceived usability with the AI app.

Results
Eighteen participants, ranging from ages 20-60, were included; 67\% (n=12) female and 33\% (n=6) male, with over 80\% having completed a higher education. The AI app outperformed the participant’s “usual course” in triage recommendations, when the expected outcome was, “go to ER/Urgent Care” (cases A and D; p values 0.015 and 0.029, respectively), but was not statistically significant in the cases with more conservative expected outcomes (B: call your provider, and C: stay home and monitor; p values 1.000, and 0.173, respectively). The time to completion—the point at which the participant was comfortable deciding when and where to seek medical treatment—using the AI app was greater, resulted in a 2.227-point increase over the participant’s “usual course” of action (Coef. = 2.227, p = 0.000, CI = 1.44 to 3.00). The correct course of action (expected outcome) was different, based on the case. When participants utilized the AI app, there was an increased chance of the participant taking a more cautious approach in seeking medical care (i.e., going to ER/Urgent Care) (odds ratio = 2.37, 95\% CI: 0.94 to 5.96, p = 0.66). The difference in level of confidence, with the decision made, was not statistically significant. (Control: 8.63 vs AI: 8.97, p = 0.250). The System Usability Scale score averaged to 85 among the participants and only 4 out of 18 participants gave a rating of 67.5 or below.

Conclusion
In this study, we found that participants would spend more time making a decision about their care when using the app, but in riskier cases, where the recommendation was to go to the ER, participants made the correct decision. Our results suggest that, while using the AI app, participants were twice as likely to choose a more cautious route of action, such as going to the ER or Urgent care. While our research findings demonstrated a noticeable increase in both time to completion as well as likelihood to seek medical attention at an ER, the level of granularity achieved through the additional questions provided by the AI seems to have the potential to increase the level of confidence experienced by users, when utilizing the application instead of a query via a search engine.
Action-oriented Artificial Intelligence for Suicide Risk Prediction: Prospective EHR-based Validation in a Large Clinical System

Michael Ripperger1, Drew Wilimitis1, William W. Stead1, Kevin Johnson1, Colin G. Walsh1
1Vanderbilt University Medical Center, Nashville, Tennessee

Introduction
Suicide prevention begins with risk identification. Numerous risk models using complex combinations of factors have been published internationally. Almost none have been implemented into clinical systems because of challenges including generalizability, data availability, reliability, transparency, and complexity. The notable exception is REACH VET, a fully-fledged prevention program with integrated predictive model. Following a five-phased framework, Action-Informed Artificial Intelligence, we completed multiple phases of validation and silent deployment within a vendor-supplied electronic health record (EHR) to predict suicide attempt risk—reaching phase five which rigorously tests system readiness for a trial of clinical decision support to enhance suicide prevention.

Methods
We analyzed prospective and concurrent validity of research-derived risk models on EHR data. Predictions were generated silently for the risk of thirty-day suicide attempt at the time of visit registration for all Vanderbilt University Medical Center (VUMC) inpatient, emergency department (ED) and ambulatory surgery encounters over ten months (6/2019-4/2020) using past diagnoses, medications, and demographics. We ascertained suicidal ideation and attempts via reference diagnostic codes. Cause-of-death data were unavailable. Statistical analyses included discrimination, e.g., number needed to screen (NNS), c-statistics, positive/negative predictive values (PPV/NPV) and Spiegelhalter’s Z-statistic for calibration. Logistic recalibration was applied using the first five months of data.

Results
The system calculated 115,905 predictions for 77,973 patients over 296 days. Performance was good across VUMC (c-statistics for ideation and attempt, respectively: 0.84 [0.84-0.84] and 0.80 [0.80-0.80]); in the Adult Hospital (0.77 [0.77-0.77] and 0.84 [0.84-0.84]); and ED (0.78 [0.78-0.78] and 0.70 [0.70-0.70]). Performance was poor in psychiatric settings (0.63 [0.63-0.64] and 0.54 [0.54-0.55]). NNS to prevent one episode of ideation and attempt in the highest predicted risk deciles were 23 and 271, respectively. NNS for attempt was 256 for men and 323 for women. NNS was 373, 176 and 407 for white, black, and non-white/non-black patients, respectively. In settings with universal screening (ED) and without Universal Screening (non-ED settings), NNS in the top risk declines were 225 and 450. Predictions were miscalibrated (S. Z statistic -3.1, p-value 0.001 across VUMC) in the first five months and improved after recalibration (S. Z statistic 1.1, p-value 0.26).

Discussion
The utility of AI for clinical practice depends on context, setting, and resource-constraints. Settings with universal screening might prioritize NPV to reduce face-to-face screening burden. Clinics with sporadic screening might prioritize PPV to direct prevention to those at-risk. Naturally high NPV and low PPV will be expected given suicidal rarity. Our study demonstrates feasible NNS in the top declines—the number who must receive a test to prevent one event. Unlike a lab test, this model relies on passively collected data and no invasive testing. Performance was best in non-psychiatric areas that resembled the research validation setting and poor in psychiatric settings where risk factors were more prevalent and patients were more likely to be treated for suicidality. Recalibration is necessary to prevent false positives and was a critical step to ensure that predictions potentially shown to providers will reflect reality.

References
Be SaskWell: Implementing a Two-Way Texting Service to Support Mental Health and Wellness Connectivity

Tracie Risling RN PhD¹, Gillian Strudwick RN PhD FAMIA², Iman Kassam MHI², Courtney Carlberg MA¹ and John Tyler Moss¹
¹University of Saskatchewan, Saskatoon, Saskatchewan, Canada; ²Centre for Addiction and Mental Health, Toronto, Ontario, Canada

Introduction

As efforts are made around the world to triage the damaging effects of the global COVID-19 pandemic, the mental health and wellness of the global population has become a critical area of focus. Loss, isolation, and a multitude of unrelenting stressors have resulted in an unparalleled mental health challenge, placing added strain and burden on overwhelmed healthcare systems and providers. The rapid development of digital health technologies over the last decade has created opportunity to employ these solutions in addressing rising mental health and wellness needs. The purpose of this study is to develop and measure the adoption of a two-way texting service, SaskWell, to improve individual awareness of and connection to existing mental health and wellness supports. As a relatively low-tech intervention with widespread reach, texting is seen as an opportunity to bridge known digital divisions in Saskatchewan communities in delivering improved access to these supports.

Methods

A three-phase mixed-methods approach was used to develop, initiate, and explore the impact of SaskWell. The RE-AIM Framework, a wide-spread implementation tool for planning and evaluating public health initiatives was used to establish three consecutive 10-week iteration cycles of texting over the course of the project. The repeated cycles allow the team to focus on the Reach, Effectiveness, and Adoption of the two-way texting service during Implementation before a concluding exploration of Maintenance. SaskWell is intended for residents of Saskatchewan 16 years of age or older who have access to a mobile device and can understand English and was developed in collaboration with a provincial patient and community advisory committee (PCAC). Users of SaskWell will be recruited through snowball sampling techniques (e.g., organizational listservs, social media, posters, news advertisements etc.). To examine the reach, effectiveness, and adoption of SaskWell, quantitative usage (i.e., engagement, click rate, response rate), polling and demographic data will be collected by users of the service and summarized through descriptive statistics and frequencies. Additionally, qualitative semi-structured interviews and focus groups with users of the service will be conducted following the end of each iteration to further explore the user experience. Interview and focus group transcripts will be analyzed using a qualitative thematic approach.

Results

SaskWell connects users with established digital mental health tools via text messaging along with cueing wellness behaviour through weekly prompts for evidenced based self-care, such as reminders of the benefits of physical activity for mental health or seeking creative outlets as positive coping mechanisms. In addition, weekly text polling questions provided additional data, including items related to the RE-AIM exploration and others tied to understanding service use and acceptability (e.g., How did you hear about SaskWell; Why did you choose to sign up; Have you used digital mental health wellness resource before?). This poster will present data from the initial phase of the project including participant recruitment numbers, initial usage data, and qualitative themes arising from both PCAC and initial user interviews. The data will provide insights into the uptake and use of service, and the benefits of delivering both low- and high-tech mental health and wellness tools via a two-way texting service.

Conclusion

As the world continues to navigate the impact of the COVID-19 pandemic, it is essential that mental health and wellness supports be accessible to all. SaskWell, a two-way texting service, is meant to connect Saskatchewan residents with tools and resources for their mental health and well-being during the COVID-19 pandemic and beyond. Building on the SaskWell foundation, the research team plans to expand this project into a national two-way texting service to support Canadian mental health and wellness as the long-terms effects of COVID-19 continue to be revealed in the months, and possibly years ahead.
Towards Measuring Real-World vs. Theoretical Impact: Evaluating Health Information Exchange (HIE) Using an Enhanced Method

Rebecca L. Rivera, PhD, MPH, CPH1,2, Heidi Hosler, BS1,2, Saurabh Rahurkar, BDS, DrPH3, Richard J. Holden, PhD, MS2,4, Joshua R. Vest, PhD, MPH1,2, Jeong Hoon Jang, PhD1,4, Jason T. Schaffer, MD4, Julia Adler-Milstein, PhD5, FACMI, Titus K. Schleyer, DMD, PhD, FACMI, FAMIA2,4

1Indiana University Richard M. Fairbanks School of Public Health, Indianapolis, IN; 2Regenstrief Institute, Inc., Indianapolis, IN; 3CATALYST Center, The Ohio State University Wexner Medical Center, Columbus, OH; 4Indiana University School of Medicine, Indianapolis, IN; 5Department of Medicine, University of California, San Francisco, CA

Introduction
At #AMIA2019 we presented the development, and #AMIA2020, the implementation, of an enhanced method to evaluate the causal effects of HIE on healthcare processes and outcomes (1,2). Combining quantitative data, such as emergency department (ED) site, HIE use logs, provider roles, visit reasons and dates with qualitative clinician survey and focus group data offers a comprehensive method to evaluate HIE implementations. It allows organizations that implement HIE across multiple sites to evaluate its outcomes in a pragmatic, meaningful way.

Objective
The objective of this study was to apply this enhanced method to evaluate the impact of implementing Health Dart, a Fast Healthcare Interoperability Resources (FHIR)-based app, on HIE use across multiple EDs within a healthcare system. Health Dart integrates clinical information related to the 7 most common ED chief complaints directly into Cerner from an HIE that was previously only accessible via web portal.

Methods
In a stepped-wedge causal study design, 14 EDs were assigned to 4 waves of 3-4 EDs based on geographic regions. This practical and robust study design allowed for each ED to serve as a control before the FHIR app was implemented between August 2019 and October 2020, thus controlling for validity threats due to historical trends (e.g. COVID-19 disruptions). HIE use, measured as FHIR app or web portal access during a patient encounter, was linked to provider and patient data. The unit of observation was the patient-level encounter clustered at the ED level (n=209,207). A logistic regression model based on a generalized estimating equations approach (GEE) was used to estimate and compare the odds of HIE use between pre- and post- app implementation, while accounting for within-ED correlation and adjusting for ED waves. Data from focus groups with clinicians examining their experience with the intervention will be conducted from April 2021 to June 2021 and presented with quantitative data in this poster at #AMIA2021.

Results
The odds of HIE use was 48% higher in ED encounters after implementation of the FHIR app than in encounters before implementation (OR=1.48; SE=0.20; 95% CI: [1.13, 1.94]; p=0.004). Among 14 EDs, the median increase in HIE use after the intervention was 2.5% (IQR: [1.9, 3.2]), with a nearly 5% increase observed for some EDs.

Conclusion
Implementation of Health Dart and an enhanced evaluation method revealed increased HIE use among ED clinicians after leveraging FHIR to integrate HIE data directly into Cerner.

References
Recommendations on Building a Sustainable AMIA Centered Around Diversity, Equity and Inclusion Practices—A Preliminary Report
Rubina Rizvi MD, PhD⁰¹; Casey Taylor Overby PhD²; C. Erwin Johnson MD, EdM,MSc³; Tiffani J. Bright PhD, FACMI¹
¹IBM Watson Health, Cambridge, MA; ²Division of General Internal Medicine, Johns Hopkins University; ³Merck CORE, Policy Evidence Research

Introduction
The topic of diversity, equity, and inclusion (DEI) was in the forefront of 2020. The initiatives to strengthen DEI-related structures and policies are relevant, now more than ever, both in the USA and globally. However, it is imperative that these DEI-related activities are long-term and sustainable, as they come with numerous challenges that are often deep-rooted, reflecting systematic and organizational problems. To further DEI efforts, AMIA must redesign the infrastructure of the systems that delay the progression and ensure sustainability of DEI within the organization. We identify and propose a list of strategic shifts and recommendations that when incorporated into DEI practices, better equip AMIA and its leaders to manage, foster, and sustain DEI.

Methods
In the context of ‘AMIA Strategic Visioning Final Report’ (SVFR), created in July of 2020, the AMIA DEI task force (TF) created three themes and corresponding subcommittees to achieve the following three targets: (1) ‘Make AMIA Diverse’—promote and build diverse representation within the AMIA membership; (2) ‘Make AMIA Inclusive’—empower AMIA members to build an inclusive professional organization; and (3) ‘Make AMIA Sustainable’—optimize the infrastructure to sustain AMIA’s DEI development and focus. We report the work of the Make AMIA Sustainable subcommittee. Each subcommittee member (COT, RR, CEJ) independently reviewed the AMIA SVFR that illuminates specific strategic and tactical suggestions based on feedback from AMIA members. Each subcommittee member created a list of strategic shifts and recommendations along with additional input from other subcommittee AMIA TF members. The Make AMIA Sustainable subcommittee met bimonthly for seven months to discuss and brainstorm ideas and learnings. All recommendations were compiled into a comprehensive document to report the goal of each proposed shift, how to implement proposed shifts, and possible outcomes. The recommendations were mapped to each subcommittee (Diverse, Inclusive, and Sustainable), along with existing AMIA committee/working groups to operationalize the recommendations. A final, full report was generated.

Table 1: Recommendations

<table>
<thead>
<tr>
<th>Major Recommendation</th>
<th>Target Strategic Area</th>
<th>Goals</th>
<th>AMIA Partners</th>
</tr>
</thead>
<tbody>
<tr>
<td>strengthen high school and undergraduate scholars with organizations that promote diversity and inclusion broadly.</td>
<td>Outreach AMIA</td>
<td>AMIA High School Relations Program</td>
<td>AMIA High School Relations Program</td>
</tr>
<tr>
<td>Visit high school, community colleges, and universities to promote diversity and inclusion broadly.</td>
<td>Outreach AMIA</td>
<td>AMIA High School Relations Program</td>
<td>AMIA High School Relations Program</td>
</tr>
<tr>
<td>Support academic and research opportunities within AMIA for faculty members in diverse backgrounds.</td>
<td>Outreach AMIA</td>
<td>AMIA High School Relations Program</td>
<td>AMIA High School Relations Program</td>
</tr>
<tr>
<td>Support academic and research opportunities within AMIA for faculty members in diverse backgrounds.</td>
<td>Outreach AMIA</td>
<td>AMIA High School Relations Program</td>
<td>AMIA High School Relations Program</td>
</tr>
<tr>
<td>Maintain and promote diversity of AMIA and the field.</td>
<td>Outreach AMIA</td>
<td>AMIA High School Relations Program</td>
<td>AMIA High School Relations Program</td>
</tr>
<tr>
<td>Promote diversity research topics and discussions in current and potential AMIA members while striving for “special focus” designation.</td>
<td>Outreach AMIA</td>
<td>AMIA High School Relations Program</td>
<td>AMIA High School Relations Program</td>
</tr>
<tr>
<td>AMIA Board/Features/Standards/Policy and Membership Outreach Academic Forum</td>
<td>AMIA Board/Features/Standards/Policy and Membership Outreach Academic Forum</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AMIA Board/Features/Standards/Policy and Membership Outreach Academic Forum</td>
<td>AMIA Board/Features/Standards/Policy and Membership Outreach Academic Forum</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AMIA Board/Features/Standards/Policy and Membership Outreach Academic Forum</td>
<td>AMIA Board/Features/Standards/Policy and Membership Outreach Academic Forum</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AMIA Board/Features/Standards/Policy and Membership Outreach Academic Forum</td>
<td>AMIA Board/Features/Standards/Policy and Membership Outreach Academic Forum</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AMIA Board/Features/Standards/Policy and Membership Outreach Academic Forum</td>
<td>AMIA Board/Features/Standards/Policy and Membership Outreach Academic Forum</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AMIA Board/Features/Standards/Policy and Membership Outreach Academic Forum</td>
<td>AMIA Board/Features/Standards/Policy and Membership Outreach Academic Forum</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AMIA Board/Features/Standards/Policy and Membership Outreach Academic Forum</td>
<td>AMIA Board/Features/Standards/Policy and Membership Outreach Academic Forum</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Results
A total of 21 unique strategic shifts were identified. Each proposed shift describes how it can be achieved, possible outcomes of each recommendation, alignment with DEI themes, and which AMIA working group can help to facilitate action. Some of the proposed shifts were applicable across more than one theme i.e., Diverse AMIA (N=8), Inclusive AMIA (N=11), and Sustainable AMIA (N=11). Key recommendations for the Governance task force (N=5) were also identified. The three major recommendations approved by the AMIA Board of Directors are listed in Table 1.

Conclusion
The Make AMIA Sustainable subcommittee comprehensively identified ways that AMIA can advance, promote, and sustain a culture of inclusion and belonging through a DEI strategy. To promote, sustain and promote a culture of DEI requires interconnected and explicit DEI goals throughout the organization.
Designing a Clinical Decision Support Framework from Primary Care Quality Performance Metric Dashboard to Individualized Patient Recommendation

Amy Robinson*, PharmD1, Tanya Podchiyska*, MS1, Samson W. Tu, MS1,2, Justin G. Chambers, BS1, Omar A. Usman, MD MBA1, Vishal Duggal, MD1, Anju Sahay, PhD1, Susana B. Martins, MD MS1, Michael Ashcraft MD1, Mary K. Goldstein, MD MS1,2 1VA Palo Alto Health Care System, Palo Alto, CA; 2Stanford University, Stanford, CA

* These authors contributed equally to this work.

Introduction
Integration of clinical quality metrics (CQM) and clinical decision support (CDS) is needed to close the quality gap between patient outcomes currently achieved in routine clinical practice and those theoretically possible when best practice treatment options are applied. Both CQM and CDS aid clinical decision making and can contribute to a unified, yet multi-faceted, quality improvement strategy. In order to provide actionable, patient-specific, evidence-based CDS for patients failing CQM, we designed a CDS framework incorporating CQM and CDS. We used CQMs to focus clinician attention on patients in need of care – those not meeting specific clinical quality targets – via CQM dashboards and actionable patient panel reports. Then we built upon our existing CDS tools to deliver the relevant assessments and clinical recommendations for bridging the quality gap and achieving optimal outcomes for these patients. The CDS provided by our system goes beyond the one-size-fits-all approach used in clinical quality measurement to incorporate individual clinical characteristics, draw upon clinical practice guidelines, and address patient complexity in multimorbidity management. Our design was informed by interviews with stakeholders, discussions with clinical advisory groups, surveys, and follow-up qualitative evaluation.

Methods
We designed a CDS framework to aid the decision-making of primary care teams in chronic disease management. The framework spans population health management and individual patient care and includes: (1) CQM dashboards to assess performance across a number of CQMs and patients, (2) drill-down reports to patient lists with actionable data, and finally (3) a patient-specific view with CQM assessments, relevant CDS clinical data, and detailed patient-specific treatment plan recommendations. The CDS is triggered by the population-based business rules of the CQM. Once triggered, our CDS tools use formalized knowledge from clinical practice guidelines and drug information resources as well as up-to-date patient electronic health record data to evaluate multiple clinical characteristics of the patient and support their multimorbidity management with individualized patient recommendations.

Results
The CQM dashboards support primary care by providing easy-to-view displays of aggregate data with respect to performance measures and may be viewed and compared at different levels: Veterans Integrated Service Network (e.g., region), medical center, clinic, and provider panel of patients. The actionable patient lists provide a population view of patients who are not meeting clinical quality targets, displaying select clinical data and information for following up with the patient. The patient-specific view allows for a more in-depth look at the patient’s clinical status and provides recommendations for clinical practice guideline-based interventions.

Conclusion
Our CDS framework enables chronic disease surveillance and management from the population to the individual patient and delivers actionable, patient-specific, evidence-based advisories for chronic disease control.

References

Views expressed are those of the authors and not necessarily of the Department of Veterans Affairs.
PROBLEM OVERVIEW

There has been a massive shift towards telemedicine given the coronavirus disease 2019 (COVID-19) pandemic with the aim to minimize viral spread, and to promote care continuity. However, it is known that older US adults experienced delays in care; therefore, to understand possible factors associated with healthcare delays, we evaluated the effect of chronic medical conditions in both telemedicine readiness and unreadiness during the COVID-19 pandemic using a nationally representative sample.

METHODS

3,379 US elders (≥ 65 years old) from the National Health Trends Study (NHATS) cohort were interviewed in 2018 prior the COVID-19 pandemic, including participants from every state except Alaska, Hawaii, and Puerto Rico, participants reported diagnosis of medical conditions. Telemedicine readiness was defined as reporting any of the following features: able to contact provider online, handle medical insurance online or getting medical information online. Reporting one of the following was indicative of telemedicine unreadiness due to a physical factor: difficulty hearing, watching TV or reading even with glasses or problem making self-understood. Lastly, technical factors leading to telemedicine unreadiness included one of the following: Not owning a telecommunication device, not being aware of how to use such device, or no recent email, text or internet use in the past month. Multivariate regression models adjusted by demographic characteristics were developed to examined telemedicine implementation across medical conditions.

RESULTS

Models were adjusted for age, gender, race/ethnicity and marital status. Telemedicine readiness was associated with cancer, hypertension, and arthritis (odds ratio [OR] ranging from 1.21 to 1.60). Most conditions were associated with telemedicine physical unreadiness (OR ranged from 1.27 to 2.48); lastly, heart disease, diabetes, and mood symptoms to technical unreadiness with OR ranging from 1.45 to 2.52.

CONCLUSIONS

Our findings highlight the importance of chronic medical conditions in telemedicine implementation. Therefore, measures to improve telemedicine implementation should involve education, guidance, and communication with healthcare workers minimizing the gap of care across medical comorbidities. As noted in participants with cancer; where the cancer continuum model has played a crucial rule in telemedicine care while improving healthcare outcomes.
Identifying Open-Source Health Information Technology Tools for Healthcare Researchers

Ronald E. Romero Barrientos1, Siddhartha Nambiar, PhD 1, Tracy Kim 1, Julia Sheehan 1, Laura C. Schubel 1, Sadaf Kazi, PhD 1,2, Kristen E. Miller, DrPH, CPPS 1,2

1MedStar Health National Center for Human Factors in Healthcare, Washington DC, USA
2Georgetown University School of Medicine, Washington, DC, USA

Introduction

Advances in provider and patient facing health information technology (IT) such as electronic health records (EHRs), and personal connected devices have contributed to a rapid explosion of healthcare data. However, the technical and research communities already face challenges related to the scale and reproducibility of these healthcare data. Therefore, it is essential to understand the current landscape in terms of challenges and opportunities related to maintaining the capacity to collect, transform and analyze these data. To help bridge this gap and to better equip researchers with accessible tools that aid in the collection, transformation, and analyzing of healthcare data, our research attempts to identify and present the current state of open-source health IT tools.

Methods

Our research team conducted a comprehensive scoping review of open-source health IT tools leveraging a modified York framework2. Librarians, technical subject matter experts (SMEs), and healthcare researchers were consulted to produce an extensive list of open-source tools that would be screened against inclusion criteria. We developed a codebook to characterize information about each tool’s licensing, documentation quality, community, longevity/pedigree, functionality, security, and support. Tools were also classified into tool types, assigned to a stage in the data lifecycle, and assessed as to whether they enabled interoperability.

Results

Abstracts (n=2,707) and tools (n=636) were retrieved and screened against the inclusion criteria. After screening, a final list of tools (n=121) were identified for domain extraction. A sample of the extracted tools (Table 1) illustrates a brief example of information collected through the scoping review. Regarding the data lifecycle, 54% (n=66) of tools involved data analysis, 31% (n=37) involved data maintenance, and 21% (n=26) involved data capture. Additionally, the tools encompassed more than 13 different open-source licenses and were categorized into 4 different tool types.

Table 1. Sample of popular open-source tools identified through our scoping review.

<table>
<thead>
<tr>
<th>Tool Name</th>
<th>Data Lifecycle Stage</th>
<th>Tool Type</th>
<th>Licensing</th>
<th>Number of Contributors</th>
<th>Age</th>
<th>Git Stars</th>
</tr>
</thead>
<tbody>
<tr>
<td>D3 (D3.js)</td>
<td>Data Analysis, Publication</td>
<td>JavaScript Library</td>
<td>BSD-3-Clause</td>
<td>52</td>
<td>2 years</td>
<td>95,500</td>
</tr>
<tr>
<td>Consent2Share</td>
<td>Data Capture, Maintenance</td>
<td>Standalone Tool</td>
<td>Apache 2.0</td>
<td>9</td>
<td>4 years</td>
<td>35</td>
</tr>
<tr>
<td>BentoML</td>
<td>Data Usage</td>
<td>ML Framework</td>
<td>Apache 2.0</td>
<td>196</td>
<td>2 years</td>
<td>2,100</td>
</tr>
<tr>
<td>Transformers</td>
<td>Data Analysis</td>
<td>Standalone Tool</td>
<td>Apache 2.0</td>
<td>3,100</td>
<td>3 years</td>
<td>39,978</td>
</tr>
</tbody>
</table>

Conclusion

The goal of our research was to explore, identify and describe open-source health IT tools to be used for research. Through our scoping review we were able to establish a repository of 121 open tools to assist the research and technical communities. By compiling, identifying, and describing a large repository of open-source tools, the research team plans to build on this research by conducting a needs assessment to gauge the gaps in the current state of open-source health IT tools. In its current state, this research may inform technical and research communities of tools that can assist them in collecting, transforming, and analyzing the abundant amount of healthcare data being produced.

References

1. Olarunke I, Oluwaseun O. Big data in healthcare: prospects, challenges and resolutions. In2016 Future Technologies Conference (FTC) 2016 Dec 6 (pp. 1152-1157). IEEE.
Evaluating Brain Age Models in Adolescents with CHD

Joy Roy, B.S. B.A.¹, Ashok Panigrahy, M.D.¹,², Rafael Ceschin, Ph.D.¹,²
¹University of Pittsburgh School of Medicine, Pittsburgh PA; ²UPMC Children’s Hospital of Pittsburgh, Pittsburgh PA

Introduction:

Term-born neonates with Congenital Heart Disease (CHD) have been shown to have developmental trajectories that are more akin to healthy preterm neonates¹. Incidentally, new research shows that adults with CHD are significantly more likely to develop early-onset dementia, suggesting accelerated aging and cognitive decline². Together, it is apparent that there is an aberrant rate of neurological development in patients with CHD during the lifespan. We will attempt to model this development and identify abnormalities through the “BrainAge” framework.

Accurate modeling of a healthy individual’s age based on structural brain imaging has the potential of providing a non-invasive biomarker of aging³. There are several approaches for modeling the normal aging process in humans, which can be used for the early identification of pathologic brain development. Over the past 10 years, the estimation of a Brain Age has gained traction as a robust and reproducible approach to this problem. These “BrainAge” models attempt to predict chronological age, or time from birth, from a large normative sample, and in doing so learn the characteristic brain morphologies that best correspond to relative ages in a general population. Subsequently, these models can be used on a sample population to estimate an individual’s deviation from a normative rate of development, measured as the difference between the predicted (phenotypic) age and the individual’s chronological age.

A significant subset of adolescents with CHD have higher incidence of executive function disorders⁴. Here, we will adapt the brain-age model to investigate the pattern of aging at the critical development period of adolescence, with the secondary goal of developing potential biomarkers predictive of executive function disorders.

Methods:

Preprocessing: Preprocessing of images is done with the use of the FMRIB Software Library (FSL) and Python 3.6. All T1-weighted images have their brains extracted, with bias field correction and tissue segmented within the same generative model (FSL FAST). The images are then registered with an affine registration to a common space. To reduce noise and search space, images are passed through a Gaussian smoothing filter and then downsampled. Principal Component Analysis (PCA) is then used to further reduce the dimensionality of the input data, with the added benefit of standardizing inputs from varying scanner sources.

Prediction: To predict the ages, we implemented a number of machine learning tools and benchmark their effectiveness independently. This study employs a supervised learning approach where we supply the models a training set of images and their corresponding chronological ages and measure their prediction accuracy with a test set. The models include Random Forest Regression, Linear Regression, and finally a 3Dimensional Convolutional Neural Network. To compare performances, we compute a Mean Absolute Error for each tool and compare them to the performance of a Relevance Vector Regression model used in the original framework³.

<table>
<thead>
<tr>
<th>Data Source</th>
<th>n Controls</th>
<th>n CHD</th>
<th>CHD Type</th>
<th>n Scan Sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single Ventricle Reconstruction III: Brain Connectome and Neurodevelopmental Outcomes</td>
<td>100</td>
<td>140</td>
<td>HLHS</td>
<td>11</td>
</tr>
<tr>
<td>Adolescent Brain Cognitive Development</td>
<td>11,878</td>
<td>-</td>
<td>-</td>
<td>21</td>
</tr>
</tbody>
</table>

References

Content Analysis and Development of a Taxonomy for Value Set Issues

Elise M. Russo, MPH,1 Arianna E. Nimocks, BSc,1 Dean F. Sittig, PhD,2
Adam Wright, PhD1

1Vanderbilt University Medical Center, Department of Biomedical Informatics, Nashville, TN; 2School of Biomedical Informatics, University of Texas Health Science Center, Houston, TX

Introduction

Value Sets are “lists of codes and corresponding terms, from National Library of Medicine (NLM)-hosted standard clinical vocabularies (such as SNOMED CT®, RxNorm, LOINC® etc.), that define clinical concepts,” that provide the foundation for clinical decision support (CDS) and quality measurement (1). The NLM Value Set Authority Center (VSAC) serves as the authoritative repository for value sets maintained by a variety of stewards, or organizations, institutions, and societies that develop and maintain value sets. Creating and maintaining value sets is challenging, and life-threatening errors in CDS or errors in eCQMs (electronic clinical quality measures) affecting payment can occur as a result of improper creation or maintenance (2). These errors are usually due to missing and/or extra codes contained in value sets.

Methods

The ONC (Office of the National Coordinator for Health Information Technology) maintains a Jira issue repository where users of value sets or systems and software that employ value sets (referred to as “reporters” here) can log, track, and discuss issues with subject matter experts (3). We queried the eCQM Issue Tracker Jira, pulling all the reported issues for which the “component” or “issue type” identified was “ValueSet” and also performed a search using the term “ValueSet” (n= 1033). We took a convenience sample of these issues (143) and then reviewed them to determine whether they were true value set errors. Of the 143 issues, 42 (29%) were true errors (non-value set errors included issues with authoring tools, browsers, or general support issues). The authors independently reviewed the reported issues and performed a preliminary content analysis using a deductive approach to determine the causes of value set errors. The individually determined error categories were discussed, adjusted, and agreed upon by all authors.

Results

Of the 42 value set errors, 32 were missing codes, 9 were additional codes and 1 was both. 18 errors pertained to SNOMED codes, 3 to ICD9, 8 to ICD10, 9 to CPT, 12 to RxNorm, 4 to LOINC, 1 to CVX, 2 to HCPCS, and 1 to CDCREC. 9 errors were associated with multiple terminologies. We identified the following causes of errors in the value sets and the number of occurrences of each:

- **Error by Steward in Value Set Creation**: The underlying guidelines and specifications were right, but someone made a simple mistake while creating the value set in the tool they were using [14 Issues].
- **Mismatch between coding systems**: A value set was written by the steward using one code system (e.g., SNOMED) but the reporter's data was coded using a different code system (e.g., ICD10) [11 issues].
- **Reporter misunderstood Steward's intent**: The reporter and value set creator simply disagree about whether a concept should be in the value set [9 Issues].
- **Clinical situation not contemplated**: A clinical situation arose that the value set creator had not contemplated, but likely would have included in the value set if he/she had [5 Issues].
- **Guideline issue**: Multiple guidelines, or multiple versions of one guideline conflict. The value set correctly reflects one of them, but the reporter expected a different one [3 Issues].

Conclusion

Categorizing errors in value sets can help us understand and tailor development of solutions to the specific problem.

References

Prediction Model for Detecting Asymptomatic Covid-19 Infection in Preoperative Patients

Amit K Saha, PhD¹, Scott Segal, MD, MHCM¹
¹Wake Forest School of Medicine, Winston Salem, NC, USA

Introduction
The COVID-19 epidemic has resulted in an ongoing worldwide epidemic with over 34 million cases in the USA and 180 million worldwide¹. Surgery on COVID-19 positive patients puts healthcare workers at risk during so-called aerosolizing procedures and also puts the patient at high risk for perioperative complications and death². At the present time, most large medical centers test all asymptomatic elective surgical patients for SARS-CoV2, and across the USA, 0.5%-2% of such patients test positive. We examined the prospect of implementing artificial intelligence to model combinations of available preoperative health records and demographic information in predicting the results of asymptomatic COVID-19 tests for patients undergoing elective surgery to derive a selective testing algorithm with low false negatives and sufficient accuracy to significantly reduce the burden of tests required. We sought to compare whether the use different sampling methods would improve predictive model performance.

Methods
With IRB approval, we conducted a retrospective observational analysis of a total of 42,377 asymptomatic patients scheduled for elective surgery at WFBMC who received a preoperative COVID-19 test from April 15, 2020 through January 14, 2021 (overall positive test rate 1.2%). No patient was tested because of suggestive symptoms or known close contact with an infected individual. Subjects were divided into a training set for model development (n= 37,128, data through December 15, 2020) and the test set (n= 5,249, data from December 15, 2020) for testing model performance. The intervals were chosen because both were prior to initiation of COVID vaccination and unchanged rules for testing. The training set was further split 80% for model building and 20% for validation. Variables included previous asymptomatic test results, demographics information, any previous hospital encounter during last year, socioeconomic characteristics, prevalence rate of COVID-19, community mobility index and comorbidities including hypertension, COPD, pulmonary disease, diabetes and cancer.

AUROC (the area under the receiver operating characteristic curve) was compared between logistic regression, random forest, and deep convolution neural network (CNN) models. The predictive model dataset consisted of 122 dimensions. Using feature reduction, we found 56 of the features were duplicative, constant or highly correlated with each other (r > 0.9). The final data set analyzed included 55 features which were standardized for predictive modelling. Because of extreme class imbalance (of all tests positive results were a very small), we used different sampling techniques (undersampling, oversampling, and cost sensitive learning) to improve model performance. Model performance validation was performed with tenfold cross validation.

Results
The preoperative asymptomatic COVID-19 positive rate for training set was 1.19% while for the testing set it was 1.2%. Based on available preoperative health records along with socioeconomic characteristics, the AUROC of the logistic regression model with cost sensitive learning (positive class weight 1:210) was 0.72, that for the CNN (positive class weight 1:270) was 0.66, and for random forest 0.71 (positive class weight 1:120). The highest sensitivity value of 98.31% with corresponding NPV of 99.86% for the logistic model with positive class weight 1:210 can be achieved with a classification cutoff at 0.33. However, to achieve this performance, 92% of all patients would still need to be tested.

Conclusion
Our single center retrospective study corroborates the concept that using machine learning algorithms, predictive models could potentially be used to predict positive COVID-19 test results for patients scheduled to have elective surgery with reasonable performance. Traditional logistic regression outperformed other machine learning algorithms. However, the best performing model still would require testing nearly the entire population, making selective testing with an acceptable false negative rate an elusive goal.

References
Characteristics of Safety Dashboards used by Harvard-Affiliated Hospitals

Hojjat Salmasian, MD MPH PhD; Michelle Frits, BA; Christine Iannaccone, MPH; Sevan Dulgarian, BS BA; Laura Myers, MD MPH; Merranda Logan, MD MPH; David M. Levine, MD MPH MA; Christopher Roy, MD MPH; Lynn A. Volk, MHS; David Shahian, MD; Elizabeth Mort, MD MPH; David W. Bates, MD MSc; Christopher Roy, MD MPH; Lynn A. Volk, MHS; David Shahian, MD; Elizabeth Mort, MD MPH; David W. Bates, MD MSc

1Brigham and Women’s Hospital, Boston, MA; 2Kaiser Permanente, Oakland, CA; 3Massachusetts’ General Hospital, Boston, MA; 4 Mass General Brigham, Somerville, MA; 5 Maine Medical Center, Portland, ME

Introduction

Healthcare organizations commonly use dashboards to track and display key performance indicators (KPIs) and metrics related to patient safety. These dashboards support clinicians and administrators in viewing outcomes of care and to drive improvement initiatives. Due to their internal and often confidential nature, these dashboards have not been subject to much research. Consequently, there is little evidence of how they compare in design and functionality, even among affiliated institutions. We analyzed the characteristics of patient safety dashboards from several affiliated hospitals to ascertain their similarities and differences in terms of design, content, and target audience.

Methods

We collected electronic copies of patient safety dashboards from a sample of ten Harvard-affiliated hospitals as part of a larger project (the Safe Care study) aimed at studying the utility of nationally recognized safety metrics for various domains of healthcare. Four reviewers independently collected the following data about each dashboard: what information is being captured, how it is presented, what platform is it present on, the timeframe for which data is displayed, frequently of data refresh, who is the target audience of the dashboard, and whether and how performance status was conveyed for each metric. Our analysis was informed by prior work in evaluation of healthcare dashboards.1-3 We also compared the dashboards within and between hospitals to identify similarities and differences.

Results

The study hospitals included 3 large academic hospitals and 7 community hospitals (including 3 with < 400 beds) which were selected to be representative of the population of Massachusetts. Hospital type and size correlated with type and number of metrics shown on the dashboards, e.g., smaller hospitals may not have dashboards for certain conditions such as hospital-acquired pressure injury (HAPI) that have a low absolute frequency in their settings. Hospitals used different platforms to produce their dashboards (Tableau, Midas, Excel, and Google Sheets). All dashboards focused on either a 2-category or 3-category approach to identify the status of each metric—either above or below a threshold level, or above, below, or within a target level. Most institutions (80%) used color coding in their dashboards to accomplish this (e.g. red/green or blue/orange), but only some (67%) used a legend to explicitly describe the coding. Only 2 dashboards indicated who was responsible for monitoring each metric. All hospitals used a mix of targets, thresholds, benchmarks, and trendlines to indicate how they were performing for each metric, and there was no consistency in their choice of which one to use across different dashboards. Most dashboards included an aim to indicate the direction the metric needed to change to improve.

Conclusion

Even among affiliated hospitals, the variability in the design and function of safety dashboards was extensive. Some sites within the same hospital group used different platforms to produce their dashboards making comparisons for similar metric trends difficult to see. Future studies should assess whether standardizing the content and display of the dashboard or use of specific dashboard design can positively influence accountability and improvement initiatives.

References

User Centered Design of a Clinical Deterioration Response System for Outpatient Cancer Patients

Megan E. Salwei, PhD, Laurie L. Novak, PhD, MHSA, Shilo Anders, PhD, Kim M. Unertl, PhD, Carrie Reale, MSN, RN-BC, Joyce Harris, MA, Jason Slagle, PhD, Leigh Anne Tang, Michelle Gomez, Zhoujun Sun, Madhavi Mani, Reena Zhang, Akhil Choudhary, Paromita Nath, PhD, Matthew Weinger, MD, Dan France, PhD
Vanderbilt University Medical Center, Nashville, Tennessee, USA

Introduction

Cancer patients are at high risk for adverse events such as medical errors and unplanned hospitalization due to the complexity and toxicity of their treatment. A substantial amount of cancer care occurs in ambulatory settings, leading to vulnerability to undetected, avoidable clinical deterioration outside of healthcare settings. A recent study found that 31% of re-admissions in cancer patients were preventable. Identifying patients at risk for clinical deterioration is a challenge as clinicians are often unaware of patient complications between visits and thus are unable to intervene in a timely manner. With the growing use of health IT, there are opportunities to leverage these technologies to identify clinical deterioration before an adverse event occurs. In this study, we describe the user centered design (UCD) process used to develop a clinical deterioration risk prediction system for cancer patients.

Methods

We are developing a predictive model that generates a patient’s risk for clinical deterioration based on FitBit®, geolocation, EHR, and patient-reported data. The patient-specific risk information will be communicated to the patient’s clinical team if the risk of deterioration becomes high. Our system will then support the clinical team in contacting the patient to prevent an unplanned treatment event (e.g., hospitalization) due to the unanticipated clinical deterioration. Following the UCD process (Figure 1), we developed a prototype of the clinical deterioration risk prediction system. To understand the environment of use (step 1), we conducted 36 observations of clinicians, staff, patients, and their caregivers, across approximately 100 patient encounters. To define user needs (step 2), we conducted interviews with 17 clinicians from a variety of clinics and with differing roles (e.g., surgeon, nurse, pharmacist). We systematically analyzed the observation and interview data to develop a mock-up of the user interface in Sketch, which was iteratively refined (Figure 2).

Figure 1: User centered design process

Results and conclusion

The observations and interviews identified design requirements for the risk prediction system, including how to integrate the system in clinic workflow. We found that nurses would be the primary users of the system, which led us to conduct additional interviews with nurses to gather their specific design feedback. Clinicians reported that our system should include the caregiver’s phone number and support messaging the patient via the patient portal. Finally, we found that patient weight changes are a significant indicator of clinical deterioration. Based on this information, we added patient weight to our system and gave patients smart scales to collect their weight via FitBit®. Next, we will gather feedback on the prototype from various clinical roles. The design and implementation of a clinical deterioration risk prediction system may prevent adverse events for outpatient cancer patients. In the design of health IT, it is important to involve diverse perspectives so that the technology adequately supports users’ work.

Acknowledgements: This research is supported by AHRQ (R18HS026616-02) and the NLM (T15LM007450-19).

References

Incorporation of American Hospital Association Annual Survey Data into a Health Services Research Data Warehouse

Nichole Samuy, MD, MSHI1, Bunyamin Ozaydin, PhD1, Ferhat Zengul, PhD1
1University of Alabama at Birmingham, Birmingham, AL, USA

Introduction
There are many challenges to developing data warehouses for health services research (1). A common obstacle is that raw datasets must undergo data preparation and integration comprising the extract, transform, and load (ETL) processes, which requires a technical proficiency that many researchers lack. The University of Alabama at Birmingham (UAB) maintains a research data warehouse (RDW) that serves as a resource for faculty and student health services researchers and includes such datasets as CMS Hospital Compare, Medicare Cost Reports, and HRSA Area Health Resources Files. Due to challenges with ETL automation, the RDW does not include American Hospital Association (AHA) Annual Survey data. To address this barrier, we developed a semi-automated process to incorporate AHA Annual Survey data and its yearly updates to the RDW. This poster will present the protocol design of data set incorporation including the process of using the incorporated data to design a multidimensional model and create data visualizations based on a hypothesis-generating research question.

Methods
The AHA Annual Survey of Hospitals is a serial cross-sectional survey that results in hospital- or system-level data involving >800 data fields from >6000 US hospitals (2). The major survey foci fall into 3 data categories and each yearly update is accompanied by an updated metadata file. After the most recent survey update data are extracted as raw CSV files from Wharton Research Data Services, a data portal with which UAB has a data use agreement, the overall incorporation/update procedures involve a four-step process (see Figure 1). Step I updates the metadata table and Steps II-IV update data category tables. Step II prepares a staging table upon which transformation occurs via the execution of a stored procedure and involves the creation of a view that is used in Step III for manual inspection of new columns. Step III also includes creating SQL statements used in Step IV to add new columns to the destination tables. Step IV finishes the process by transforming the new data to the correct data types, importing them to the destination tables, and dropping staging tables. Data quality issues such as missing values were identified by manual inspection after running verification queries, and SQL statements were added to the Step II SSIS package to address the presence of NULL values. After data incorporation, we created a multidimensional data mart including a star schema with measures representing physician and nurse staffing ratios and dimensions including year, geographic location, hospital type, hospital ownership, and bed size.

Results and Conclusion
In PowerBI, we created multiple visualizations including staffing ratios by state, over time, by bed size, and by primary service. To demonstrate the power of our designed process, we linked data from AHA Annual Survey to CMS Hospital Compare based on a common unique hospital identifier. Specifically, we evaluated Hospital Compare measures of stroke education and preventable venous thromboembolism by calculated AHA Annual Survey nurse staffing ratios and generated research hypotheses that could lend themselves to more rigorous study design and statistical analysis.

In conclusion, we successfully created a reproducible, semi-automated process to accommodate integration of AHA Annual Survey data to the RDW that could be applied to SQL server-based data warehouses at any institution with access to AHA Annual Survey data. Finally, we were able to demonstrate the value of this process by linking to other datasets in order to address research questions that could not be answered with a single data set alone.

References
Analyzing Associated Data Tracking via PubMed Secondary Source ID

Amanda R. Sawyer, MLIS, Vojtech Huser, MD PhD
National Library of Medicine, Bethesda, MD

Introduction
The semantic web promised a revolution in making web content machine readable. For healthcare research, semantic publishing offers opportunities to link literature with associated datasets with the goal of promoting increased transparency and reproducibility. Structured article metadata in PubMed represents partial fulfillment of this vision. For example, the Secondary Source ID (SI) field in PubMed tracks associated data found in life science databases (e.g., SWISSPROT) and clinical trial registries (e.g., ClinicalTrials.gov). Prior to 2016, publishers’ ability to edit databases in the SI field was limited. MEDLINE indexers reviewed each citation and added relevant databases from a controlled list of sources. Due to the ever-increasing volume of research, manual review became untenable and new policy increased journal involvement in article metadata management. Since 2016, publishers have been tasked with including external database sources in the SI field of the article XML submission, MEDLINE indexers manually review only a minority (around 45%) of citations. This study examines the content and past and future added value of SI article metadata.

Methods
The PubMed SI field currently tracks 42 databases, including 20 clinical trial registries and 22 biology databases such as molecular sequence databases or data sharing repositories. We examined the frequency these databases appeared in article SI metadata in a recent year (2019) and analyzed the trend over time by year (2000–2019). Given increased involvement of publishers in article metadata following 2016, we paid particular attention to how this change impacted the trends of individual databases by comparing SI metadata from 2016 to 2017. This analysis was motivated by another project that linked journal articles reporting results of observational research to healthcare real-world data (RWD) databases (e.g., Clinical Practice Research Datalink [CPRD] or PEDSNet) that were utilized in the article.

Results
In 2019, nearly 1.4 million citations were added to PubMed. Of those, 23,662 (1.7%) citations included a value in the SI field. ClinicalTrials.gov and GENBANK were among the most frequently cited databases. Some registries and databases included in the tracked list such as the Peruvian Clinical Trial Registry, BioProject, and the PubChem Compound Database were rarely or never cited. Trend over time analysis revealed some substantial changes to database frequency in SI metadata from 2016-2017. Several databases have declined in citations, likely due to the metadata authoring policy changes. For example, 12 of the 20 tracked clinical trial registries decreased in citations by at least 20%. 10 of the 22 tracked biology databases decreased in citations by at least 20%. Databases identified as priorities for MEDLINE indexers increased in frequency in the SI field, such as figshare (177.5% increase) and Dryad (6% increase). Complete trend data and additional results can be found online in the GitHub repository for this project.

Conclusion
The motivation for this study stemmed from an interest in evaluating RWD databases as candidates for SI metadata. However, given that publishers are the primary authors of metadata and citations including SI metadata have decreased for many databases in the last several years, determining an appropriate scope for the SI field presents an important semantic challenge.

References
Utilizing a Clinical Decision Support Alert to Increase Nasal MRSA PCR Ordering and Reduce Vancomycin Course Duration

Diana J. Schreier, PharmD, MBA, Patrick M. Wieruszewski, PharmD

1Mayo Clinic, Rochester, MN

**Background:** Despite widespread availability of nasal MRSA tests, a quality improvement project revealed a lack of knowledge amongst clinicians regarding the utility of the test in de-escalation practices. This resulted in low rates of test ordering and potential missed opportunities for vancomycin de-escalation, contributing to potentially prolonged antimicrobial exposure, increased opportunity for vancomycin-induced drug toxicity, and excess resource utilization.

**Methods:** The first iteration of the quality improvement project occurred from September 3, 2018 to September 2, 2019 and involved the provision of education on the utility of nasal MRSA PCR tests and authorized pharmacists to order the test if the prescriber did not order it when placing the vancomycin order. The second iteration of the quality improvement project occurred from September 3, 2019 to September 2, 2020 and included the implementation of a clinical decision support alert in the pharmacist verification queue when a prescriber placed a vancomycin order with a respiratory infection indication, but a nasal MRSA PCR had not been ordered or resulted in the past 7 days (Figure 1).

**Results:** Of 2,183 eligible patient encounters that occurred during the first iteration, MRSA nasal PCR tests were ordered 29.9% (n=652/2,183) of the time. This rate increased to 67.4% (n=1,573/2,331) after the implementation of the clinical decision support alert in the second iteration (Figure 2). The increase in nasal MRSA PCR testing was also correlated with a decrease in median empiric vancomycin course durations from 33.8 (12.0, 54.5) to 27.5 (12.0, 47.8) hours.

**Conclusion:** Clinical decision support tools can be used to reduce vancomycin overuse and improve adherence to antimicrobial stewardship initiatives.

![Figure 1. Clinical decision support alert display in the pharmacist verification queue when an order for vancomycin with a respiratory infection indication is placed.](image)

![Figure 2. Percent of unique patient encounters that met clinical decision support alert criteria and a MRSA nasal swab was ordered within 6 hours of vancomycin initiation.](image)
Using Mobile Integrated Health and Telehealth to Support Transitions of Care among Heart Failure Patients; MIGHTy Heart study protocol

Leah Shafran Topaz PT, MSc1, Brock Daniels MD, MPH1, Kevin Munjal MD, MPH, MSCR2, Meghan Reading Turchioe PhD, MPH, RN1, Rainu Kaushal MD, MPH1 Ruth Masterson Creber PhD, MSc, RN1

1 Weill Cornell Medicine, NewYork-Presbyterian, Department of Population Health Sciences, NYC, NY, 2Emergency Medicine / Population Health, Icahn School of Medicine at Mount Sinai, NYC, NY

Introduction Heart failure (HF) is a cardiac condition that results in a substantial burden on both health systems and patients because of preventable hospital admissions. The all-cause 30-day readmission rates for HF are 20-25% among Medicare patients, the highest for any single condition in the U.S. Outpatient follow up programs and telehealth interventions have been created to prevent readmission among this population, however, due to low appointment availability, challenges surrounding transportation and inability to administer therapeutic interventions at patients homes, these interventions have not consistently demonstrated improvement in clinical outcomes. Mobile integrated health (MIH) is a clinical informatics intervention including telemedicine and community paramedicine. The purpose of the, “Using Mobile Integrated Health and Telehealth to Support Transitions of Care among Heart Failure Patients” (MIGHTy Heart) study is to evaluate the effectiveness of an MIH intervention compared a transitions of care intervention on health utilization, patient reported outcomes and healthcare quality. We also aim to evaluate the factors that support the adoption and implementation of an MIH program from the perspective of multiple key stakeholders (e.g., nurses, paramedics, physicians, patients etc.).

Methods/Results The MIGHTy Heart study is a pragmatic randomized clinical trial (CT.gov # NCT0466254). The dual primary outcomes include hospital readmissions (measured using Medicare claims data at 30 days) and change in quality of life (measured using the Kansas City Cardiomyopathy Questionnaire). Secondary outcomes include preventable emergency department visits, unplanned hospital admissions from the emergency department at 30, 60 and 90 days (using a combination of Medicare claims, the electronic health record); and changes in PROMIS-29 outcome assessments at 30 days. Participants are prospectively enrolled during hospitalization across two health systems in NYC and randomized to either the MIH (intervention) or the transitions of care (comparator) intervention arms (Fig. 1). REDCap is used for all data collection. Follow-up surveys are sent to participants electronically and followed up by phone if not completed in time. The MIH group receives access to specially trained community paramedics who provide medical assessment and care at patients’ home under the guidance of remote emergency department physicians. During the MIH encounter, the medics use a standardized HF assessment, and an emergency medicine physician is contacted via telehealth. Physicians can access the patients’ chart via the institutional EHR to adjust to medications. Diagnostic testing, such as ECG and point-of-care labs, can be performed and treatments such as intravenous diuretics can be administered by paramedics. Participants in the comparator arm receive a follow-up phone call within 48-72 hours of discharge to evaluate their clinical and social status, identify unmet needs, reinforce patient education and confirm follow up appointments. In addition, we will conduct a multi-method study of the factors supporting the adoption, implementation, and maintenance of MIH within the two health systems, including qualitative interviews with key stakeholders and quantitative evaluation of specific implementation parameters, using the Reach Effectiveness Adoption Implementation and Maintenance (RE-AIM) framework. To date, 141 participants have been enrolled, 46.2% female (n=55), mean age 73.13 ± 14.63 years. Trial recruitment is ongoing.

Conclusion Overall, the MIGHTy Heart study will provide evidence regarding the comparative effectiveness of the MIH clinical informatics intervention for patients with heart failure after a hospitalization.
Impact of Social Determinants of Health on Predictive Models in 30-Day Hospital Readmission or Death for Patients with Severe Obesity

Marianne Sharko1, Yongkang Zhang1, Yiye Zhang1, Evan Sholle1, Sajjad Abedian1, Meghan Reading Turchhoe1, Jessica S Ancker2

1. Weill Cornell Medical College, New York, NY; 2. Vanderbilt University Medical Center, Nashville TN

Introduction: Predictive models identify patients at higher risk of readmission or death to provide information for targeted interventions. Existing models, such as the HOSPITAL score model,(1) usually do not incorporate social determinants of health (SDH) information, although this is important to health disparities. Our previous study found that the incorporation of SDH risk factors into the HOSPITAL score model resulted in a small improvement in risk prediction. This improvement was the largest for vulnerable patient groups, including those with obesity. In this study, we hypothesized that the impact of SDH on readmission risk prediction would increase with the severity of obesity.

Methods: We extracted EHR admission data between January 2015 and November 2017 in an academic medical center in New York City, including individual SDH (e.g., race/ethnicity, primary language) and patient residential zip-codes to link to census tract SDH (e.g., median income level, walkability score). We applied the simplified HOSPITAL score model to predict avoidable 30-day readmissions or death (model a). We augmented the model with three sets of predictors: individual SDH (b); census tract SDH (c); individual plus census tract SDH (d). We ran these models on patients with different classes of obesity and compared C-statistics for the performances.

Results: Our study included 14,439 admissions, 26.2% with obesity. The HOSPITAL score for non-obese patients produced a C-statistic of 0.652. Adding individual and census-tract SDH produced significant but small improvements for non-obese patients and those with class 1 obesity. However, there were greater improvements for those with class 2 and 3 obesity. The largest significant increase was found for those with class 3 obesity, with an increase from 0.679 to 0.761. Logistic regression identified significant risk factors for readmission and death based on level of obesity: class 1, dual insurance; class 2, non-English language, foreign born, lower walkability score, and class 3, single marital status, female sex, lower median house income, and lack of insurance.

Conclusions: Predictive modeling to identify patients with severe obesity who are at risk for readmission or death is augmented by the inclusion of individual and census tract SDH risk factors.

C-statistics for HOSPITAL Score Readmission Risk Predictions

<table>
<thead>
<tr>
<th>Patients Body Mass Index (BMI)</th>
<th>n</th>
<th>Hospital Score (a)</th>
<th>Hospital Score + Individual SDH (b)</th>
<th>Hospital Score + Census tract SDH (c)</th>
<th>Hospital Score + Individual SDH + Census tract SDH (d)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-obese (BMI&lt;30)</td>
<td>10,663</td>
<td>0.6520</td>
<td>0.6598 p=0.001</td>
<td>0.6544 p=0.052</td>
<td>0.6605 p=0.001</td>
</tr>
<tr>
<td>Class 1 (BMI 30&lt;35)</td>
<td>2,004</td>
<td>0.644</td>
<td>0.6636 p=0.041</td>
<td>0.6570 p=0.061</td>
<td>0.6722 p=0.012</td>
</tr>
<tr>
<td>Class 2 (BMI 35&lt;40)</td>
<td>939</td>
<td>0.6818</td>
<td>0.7098 p=0.043</td>
<td>0.7088 p=0.056</td>
<td>0.7314 p=0.004</td>
</tr>
<tr>
<td>Class 3 (BMI &gt;=40)</td>
<td>833</td>
<td>0.6792</td>
<td>0.7201 p=0.010</td>
<td>0.7330 p=0.004</td>
<td>0.7608 p=0.001</td>
</tr>
</tbody>
</table>

Notes: p values indicate statistical significance of differences in C-statistics between each of the columns (b)-(d) and column (a).

This study was supported by: Weill Cornell Medicine Dean’s Diversity and Healthcare Disparities Research Award (Co-PIs: Ancker and Yiye Zhang)

References


A multi-label classifier to screen different types of substance misuse in hospitalized patients

Brihat Sharma, MS1,2, Dmitrii Dligach, PhD2, Hale M. Thomson, PhD4, Matthew M Churpek, MD, MPH, PhD3, Niranjan S. Karnik, MD, PhD1, Majid Afshar, MD, MSCR3

1Department of Psychiatry and Behavioral Science, Rush University Medical Center, Chicago IL; 2Department of Computer Science, Loyola University Chicago, Chicago IL; 3Department of Medicine, University of Wisconsin-Madison, Madison WI

Introduction: Manual screening with self-report questionnaires is the predominant modality to identify substance misuse in hospitalized patients; however, these tools lead to poor adherence because they require additional time and staffing outside of usual care. Our group previously built discrete classifiers for detecting alcohol and opioid misuse to overcome the barriers to manual screening.1-2 We aim to further improve efficiency by building a single multi-label neural network-based classifier that can identify alcohol and opioid misuse using admission notes collected during the usual care from the electronic health record (EHR).

Methods: The training data (n=49530) and validation data (n=5504) were derived from adult hospitalized patients at Rush University Medical Center (RUMC) between October 2017 and December 2019. The Alcohol Use Disorder Identification Test questionnaire and the Drug Abuse Screening Test questionnaire served as the reference labels. The test data were composed of adult patients hospitalized at a separate health system, Loyola University Medical Center (LUMC). The separate health setting at LUMC for testing served as a more rigorous external validation. The reference labels at LUMC for alcohol misuse (n=999) and opioid misuse (n=992) were identified by expert annotation of the patient encounter in the EHR. We pre-processed the first 24 hours of EHR notes using the Clinical Text Analysis and Knowledge Extraction System (eTAKES) to map the text to concept unique identifiers (CUIs) from the UMLS metathesaurus. We used the CUIs as inputs to our machine learning model. A random search approach was applied to optimize the hyperparameters on the validation data tuned to the precision-recall area under the curve (PR AUC). Our final model was a convolutional neural network (CNN) with 1,024 filters of size 5 that consisted of an embedding layer followed by a convolution, fully-connected layers of 32 units, and three sigmoid outputs for alcohol misuse, opioid misuse, and other drug misuse (amphetamine, cocaine, and non-prescribed benzodiazepines). Results are reported on the LUMC test dataset for alcohol and opioid misuse with test characteristics and 95% Confidence Intervals (CI) for the optimal cutpoint derived from Youden’s J statistic.

Results: Our training data corpus at RUMC had 664,836 clinical notes and 37,317 unique CUIs. The case-rate for substance misuse was 4% (n=2,338) at RUMC. The PR AUC of the CNN classifier and the associated test characteristics on the test dataset at LUMC for alcohol and opioid misuse are listed in Table 1. The calibration slope and intercept for alcohol misuse were 0.81 (95% CI 0.71-0.92) and 0.29 (95% CI 0.12-0.46), respectively. For opioid misuse, they were 0.72 (95% CI 0.62-0.82) and 0.08 (95% CI -0.13-0.29), respectively.

Table 1. Results of external validation of a multi-label substance misuse classifier

<table>
<thead>
<tr>
<th>Multi-label CNN Classifier</th>
<th>PR AUC (95% CI)</th>
<th>ROC AUC (95% CI)</th>
<th>F1</th>
<th>PPV/Precision (95% CI)</th>
<th>Sensitivity/Recall (95% CI)</th>
<th>NPV (95% CI)</th>
<th>Specificity (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alcohol Misuse</td>
<td>0.92 (0.90, 0.94)</td>
<td>0.89 (0.87, 0.91)</td>
<td>0.85</td>
<td>0.84 (0.81, 0.87)</td>
<td>0.85 (0.82, 0.88)</td>
<td>0.79 (0.75, 0.83)</td>
<td>0.78 (0.73, 0.81)</td>
</tr>
<tr>
<td>Opioid Misuse</td>
<td>0.90 (0.87, 0.92)</td>
<td>0.92 (0.90, 0.94)</td>
<td>0.81</td>
<td>0.82 (0.77, 0.90)</td>
<td>0.75 (0.63, 0.85)</td>
<td>0.90 (0.87, 0.92)</td>
<td>0.91 (0.88, 0.95)</td>
</tr>
</tbody>
</table>

Note: Abbreviations (ROC AUC - Receiver Operating Characteristics Area Under the Curve, PPV - positive precision value, NPV- negative precision value)

Discussion: Our multilabel substance misuse classifier had good discrimination in our external validation dataset. An advantage of using a joint model includes information sharing across different types of substance misuse within the model, potential gains in efficiency, and the flexibility it provides to make individual substance misuse predictions. Our next steps include using multi-task learning and pre-training methods to further improve calibration. Our research is supported by the National Institute of Health (R01DA051464).

References:
Radiographic Manifestations in Tuberculosis Patients Coinfected with Diabetes Mellitus: A Systematic Review and Meta-Analysis

Lincoln R. Sheets, MD, PhD, FAMIA1-2, Leonardo Martinez, PhD3, Trang Ho Thu Quach, MPH4, Yang Ge, MPH4, Rosina Marchan, BS5, Philip Hill, PhD6, Henok G. Woldu, PhD2
1School of Medicine, University of Missouri, Columbia, Missouri; 2Center for Health Analytics for National and Global Equity, Columbia, Missouri; 3School of Public Health, Boston University, Boston, Massachusetts; 4College of Public Health, University of Georgia, Atlanta, Georgia; 5School of Public Health, The University of Texas at Arlington, Arlington, Texas; 6Otago Global Health Institute, Otago, New Zealand

Introduction
Studies show inconsistent associations of diabetes mellitus (DM) with tuberculosis findings1, which could increase misdiagnosis2. Our objective was to pool published radiographic findings in tuberculosis with and without diabetes.

Methods
We systematically searched PubMed, EMBASE, Web of Science, and BIOSIS for studies of radiographic findings in tuberculosis patients with or without diabetes, and calculated the odds ratios of cavitary disease, lower-lung infiltrates, upper-lung infiltrates, consolidation, bilateral disease, and pleural effusion.

Results
Fifty-eight unique studies presented radiographic findings for 61,903 tuberculosis patients, of whom 12,655 had comorbid diabetes mellitus. Table 1 shows pooled sample sizes and the odds ratios of each finding in diabetic patients.

Table 1. Pooled odds ratios (OR) and confidence intervals (CI) of radiographic findings in tuberculosis patients

<table>
<thead>
<tr>
<th>Radiographic Finding</th>
<th>Study Count</th>
<th>Pooled Sample (and %DM)</th>
<th>OR (and 95% CI), *p&lt;0.05</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cavitation</td>
<td>56</td>
<td>61,431 (21%)</td>
<td>1.67 (1.48, 1.87)*</td>
</tr>
<tr>
<td>Lower-lung infiltrates</td>
<td>17</td>
<td>6,093 (43%)</td>
<td>1.61 (1.24, 2.10)*</td>
</tr>
<tr>
<td>Upper-lung infiltrates</td>
<td>17</td>
<td>7,134 (36%)</td>
<td>0.67 (0.44, 1.01)</td>
</tr>
<tr>
<td>Consolidation</td>
<td>13</td>
<td>2,671 (34%)</td>
<td>1.43 (1.02, 2.00)*</td>
</tr>
<tr>
<td>Bilateral disease</td>
<td>12</td>
<td>6,119 (33%)</td>
<td>1.32 (1.06, 1.66)*</td>
</tr>
<tr>
<td>Pleural effusion</td>
<td>17</td>
<td>5,288 (38%)</td>
<td>1.33 (0.90, 1.98)</td>
</tr>
</tbody>
</table>

Conclusion
Meta-analysis found higher odds of cavitary, lower-lung infiltrates, lung consolidation, and bilateral disease in tuberculosis patients with diabetes, but no higher odds of pleural effusion or upper-lung infiltrates. Because diabetics have a higher risk of active tuberculosis3, this comorbidity may impact future tuberculosis control efforts.

References
Content in Context – Individualizing Knowledge Resources for Delivery at the Point of Care

Jane L. Shellum, M.H.A.¹, Robert R. Freimuth, Ph.D.¹, Davide Sottara, PhD¹, Aaron Nathan, MS¹, Adam Bartscher ¹
¹Mayo Clinic, Rochester, MN

Introduction
We have previously described the AskMayoExpert (AME) application (1), which was developed to provide answers to clinical questions at the point of care. AME is based on the concept of medical “gist” – concise, relevant, and clinically-applicable answers to clinical questions, written, vetted, and reviewed by Mayo Clinic faculty. The tool is infobutton-enabled and provides a pass-through search to complementary online knowledge resources, offering a single starting point for knowledge retrieval. While AME contains highly curated medical knowledge and is widely adopted by providers, it lacks the ability to situate knowledge within the context of a particular patient. In order to reduce the cognitive effort of clinicians using the application, we propose an approach based on the individualization of the knowledge content, using patient-specific and other contextual data collected at the point of decision-making.

Methodology
Knowledge content is authored using the Darwin Information Typing Architecture (DITA), an OASIS standard (2) for structured documents. DITA semantic markup facilitates the reuse of content and supports conditional formatting through attributes that allow applications to show, hide, and/or (de)emphasize fragments of content based on patient data, including lab values and observations, and keys that allow the application to insert patient-specific values into the content. The Patient data is retrieved using standard FHIR enterprise APIs; a production rule engine processes the data and generates the attributes and key values. A dedicated individualization engine applies the dynamic values to the static content and a rendering engine transforms the resulting DITA XML document into the HTML. The HTML is delivered via a standard SMART-on-FHIR application built on top of a ReSTful Service Oriented Architecture. The application is called using CDS hooks and runs inside of the EHR for a seamless user experience.

Figure 1. The Content Individualization Architecture

Results and Discussion
A usability study of 23 clinicians found that the individualized content was well received. Scalability is a concern; existing Word and HTML content could be migrated to DITA, but a non-trivial amount of editing was necessary to refactor the content and add metadata. To work in real time, efficient individualization and rendering engines had to be developed in house. The content was designed to revert gracefully to the original, static content when patient data required for individualization was not available. In this work, we have demonstrated one of many possible delivery mechanisms; there are many other possible integration patterns that could be explored.

Table 1. Static vs Individualized Content (Example)

<table>
<thead>
<tr>
<th>Static Content</th>
<th>Individualized Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medications: A, B, C, D</td>
<td>Currently on 3 medications: A, B, D</td>
</tr>
<tr>
<td>If BP is not controlled on 1 medication, …</td>
<td>If BP is not controlled on 1 medication, …</td>
</tr>
<tr>
<td>If BP is not controlled on 2 medications, …</td>
<td>If BP is not controlled on 2 medications, …</td>
</tr>
<tr>
<td>If not at BP goal using 3 medications, consider …</td>
<td>BP of 138/94 is not at goal, consider Referral</td>
</tr>
</tbody>
</table>

Conclusion
While infobutton queries can leverage some context to find applicable knowledge resources, the context is not applied to the content itself. The conditional formatting capabilities of DITA can be used to individualize content to evolve a knowledge resource into a context-aware, cognitive support application. Future work will include evaluation of user experience and cognitive burden. We will also compare its effectiveness to more traditional forms of clinical decision support and consider options for integration with clinical workflows.

References
2. Darwin Information Typing Architecture.
   http://docs.oasis-open.org/dita/dita/v1.3/dita-v1.3-part3-all-inclusive.html
Digital Mental Health Priorities of People Affected by Mental Health Problems: A Pan-Canadian Survey

Nelson Shen, PhD1, Iman Kassam, MHI1, Navi Boparai, MHA1, Clement Ma, PhD1,2, Sheng Chen, PhD1,2, Wei Wang, PhD1,2, Damian Jankowicz, PhD1, Gillian Strudwick, PhD1,2
1Centre for Addiction and Mental Health, Toronto, Ontario, Canada; 2University of Toronto, Toronto, Ontario, Canada;

Introduction

The COVID-19 pandemic has accelerated the much-needed innovation and scaling of digital mental health (DMH). Engaging people affected by mental health problems to understand their experiences and perspectives is critical in ensuring that DMH initiatives provide care that is aligned with population needs.1 A pan-Canadian survey was conducted to understand the priority areas in which digital mental health tools can support people affected by mental health problems (i.e., people who are seeking care, currently receiving care, or care partners).

Methods

An online survey was administered in February 2021 by a research technology and data collection company to their national web-based patient pool. A proportional quota sampling based on age, gender, and geographical region was used to capture a sample proportional to the Canadian population. The survey consisted of discrete-choice questions on their digital mental health experiences and priorities. Open-ended questions were used to allow participants to expand on their responses. Descriptive statistics were calculated, and a qualitative content analysis was conducted.

Results

The online survey was completed by 1003 Canadians (410 seeking care, 271 currently receiving care, and 322 care partners) and had a response rate of 27%. Three-quarters of participants (n=766) have used one or more DMH tools to support their care. The top 5 uses were: information searching (n=524); wellness (n=388); virtual care (n=340); care coordination (n=335); and activity/symptom tracking (n=235). A minority of participants were satisfied with the availability of DMH tools (n=287). The remainder of users were either unsatisfied (n=117) or unsure (n=362). Privacy/security concerns (n=331), awareness of tools (n=327), costs of tools (n=221) and equipment (n=161), and aversion to using DMH tools (n=213) were the top barriers to access. The most common open-ended reasons (n=241) for aversion to DMH tools included: need for human interaction (n=83); lack of trustworthiness (n=59); not useful or irrelevant (n=31); questions about effectiveness (n=21); and lack of awareness of tools (n=9). Collectively, the top 5 priorities for digital mental health tools to support themselves or loved ones are: DMH tool curation resource (n=590); mental healthcare system patient navigation tool (n=541); DMH record passport (n=479); secure provider messaging (n=471); and DMH tool hub/library (n=389).

Conclusions

A majority of Canadians have used some form of DMH tool to support themselves or their care partners; however, many were unsatisfied with current offerings or were unsure about what was available to them. Most non-users did not want to replace the human element of care, while others questioned the trustworthiness, relevance/usefulness, and effectiveness of the tools. These factors may add additional burden to those who are navigating the complexities of the digital and physical mental healthcare landscapes1,2. As such, DMH tools that help navigate both spaces were identified as the top priorities. Understanding the perspectives of those affected by mental health problems will allow organizations to focus their efforts to address areas of need and develop supports to foster greater understanding, awareness, access, acceptance, and adoption of DMH tools.

References

Accessing Electronic Medical Record Data Through a Breast Cancer Prevention Decision Support Tool: Usability and Pilot Testing of a SMART on FHIR API

Thomas Silverman MPH¹, Jacquelyn Amenta², Jill Dimond PhD³, Cynthia Law MPH⁴, Tarsha Jones PhD⁵, Katherine D. Crew MD, MS⁵, Rita Kukafka DrPH, MA¹
1: Columbia University Department of Biomedical Informatics, New York, NY, USA 2: Columbia University Mailman School of Public Health, New York, NY, USA 3: Sassafras Tech Collective, Ann Arbor, MI, USA 4: Columbia University Department of Medicine, New York, NY, USA 5: Florida Atlantic University, Boca Raton, FL, USA

Abstract: We conducted three usability sessions on a decision support tool’s SMART on FHIR API. After modifying the tool based on the usability session findings, we pilot tested the tool at multiple clinics.

Introduction:
We developed a decision support tool called RealRisks that allows users to access their medical record data using a SMART (Substitutable Medical Applications, Reusable Technologies) on FHIR (Fast Healthcare Interoperability Resource) supported application programming interface (API). We conducted three usability studies on this functionality and pilot tested the tool at multiple healthcare organizations.

Methods:
We invited patients at Columbia University’s Irving Medical Center to participate in virtual usability sessions where they completed a prototype of the RealRisks module that uses the API. We recruited participants by email from a list of prior research participants who agreed to be contacted for future research. We conducted the usability sessions using Zoom. Participants shared their screens while navigating the module and expressing their thoughts aloud. We prompted participants with questions on their understanding of the process and aims of the module as well as the data the API returned to them. Sessions were recorded, and the recordings were reviewed and summarized by two researchers (TS and JA). After modifying the module based on the usability study findings, we piloted the module with patient advocates at clinical sites participating in a multisite trial aiming to evaluate RealRisks’ effectiveness in promoting informed choice for breast cancer chemoprevention.

Results:
Usability successes: All three participants understood the purpose of the API module and were able to easily navigate through it. All three participants were able to access their medical record data using RealRisks, confirmed that the data RealRisks pulled and presented was accurate, and were able to edit fields when necessary.

Usability issues: One participant did not realize she had been redirected to her patient portal and needed to log in using her patient portal credentials. All three participants were confused by the language in the consent form template provided by their patient portal (open.epic). For example, participants did not understand that the consent form was referring to RealRisks when using the phrase “app developer.” All three participants had trouble understanding the medical terminology of the data that was returned to them.

Response to usability issues: We created a “what to expect” section to help navigate the patient through the module and added more descriptive text to the data labels that are provided when a user is presented her medical record data. Where we could, we edited the text in the patient portal consent form to make it more descriptive.

Multisite pilot testing: Patient advocates at eight healthcare organizations across the US were able to access their EHR data using the API module. All of these organizations use Epic EHRs for their medical record/patient portal.

Conclusion:
The RealRisks SMART on FHIR API successfully allowed users to access their medical record data. Participants found the usability of this functionality to be acceptable, but they requested detailed descriptions of the process and data involved. After modifying the tool based on these requests, we successfully piloted RealRisks at multiple healthcare organizations that use Epic EHRs. More development is needed to support the API’s functionality with non-Epic products.
Integrative Analyses Identify Susceptibility Genes Underlying COVID-19 Hospitalization

Kritika Singh*1, Gita A. Pathak*2, Tyne W Miller-Fleming1, Frank R Wendt2, Nava Ehsan3, Kangcheng Hou4, Ruth Johnson4, Zeyun Lu5, Shyamalika Gopalan5, Loic Yengo6, Pejman Mohammadi3, Bogdan Pasaniiuc6, Renato Polimanti3, Lea K Davis4, Nicholas Mancuso5

*Equal contribution

1 Vanderbilt University Medical Center, Nashville, TN, USA
2 Yale School of Medicine, New Haven, CT USA
3 The Scripps Research Institute, La Jolla, CA, USA.
4 University of California Los Angeles, Los Angeles, CA USA
5 University of Southern California, Los Angeles, CA USA
6 The University of Queensland, Brisbane, Australia

Coronavirus disease 2019 (COVID-19) is caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) which rapidly progressed into a global pandemic. 10-20% of patients known to be infected with SARS-CoV-2 need hospitalization, and among them, a fraction face significant morbidity and mortality. The host’s genetic background is likely to contribute in explaining such diverse clinical outcomes. While previous efforts have demonstrated the role of ACE2 and TMPRSS2 in host defense against COVID-19, there remains limited understanding for the role of host genetics contributing to severe COVID-19 outcome variability. Our work integrates mRNA expression, splicing, and protein abundance data with GWAS of COVID-19 related hospitalization (n=7,885 cases, 961,804 controls; Freeze 4 COVID-19 HGI excluding 23andMe participants*1) to map genes and pathways involved in COVID-19 severity. We performed mRNA/splicing/protein transcriptome-wide association studies (TWAS/spTWAS/PWAS) to identify genes across different genomic regions whose genetically predicted activity is associated with COVID-19 related hospitalization. We investigated the functional role of these genes using phenotype-wide (PhesWAS) and laboratory-wide (LabWAS) association scans to map the functional consequences of these genes in European and African ancestry patients from the Vanderbilt University Medical Center biobank (BioVU; n=85,460). We replicated phenotypes identified from BioVU in various secondary cohorts of multi-ethnic individuals. Using the predicted expression of 22,207 genes across 49 tissues for association with COVID-19 related hospitalization, we identified 123 associations representing 21 genes across 45 tissues at 8 independent genomic regions (p < 2.3E-6). Next, we tested 131,376 splice sites of predicted alternative-splicing expression identified 420, associations representing 43 splice variants for 11 genes across 49 tissues and 5 genomic regions. Comparing genes identified from TWAS (13 genes) and spTWAS (11 genes), 9 genes were implicated by both approaches. Lastly, PWAS using 1,031 predictive models of plasma proteins fitted from population data in the INTERVAL study showed significant associations with the ABO and OAS1 genes. We observed that associated phenotypes and gene functions converged on cytokine-cytokine receptor signaling involved in inflammatory response and on JAK-STAT signaling pathways involved in antiviral host response (p < 0.05). We investigated the clinical influence of these genes by performing a PhesWAS and a LabWAS on their predicted expression values. For the PhesWAS, overall, 40 clinical phenotypes were significantly associated with genetically-predicted ABO, IFNAR2, and CCR1 expression levels; ABO accounted for the majority (30 out of 40) of the associations. Across the 17 phenotype categories, we found circulatory system-related phenotypes were enriched for association (7.23-fold enrichment, p = 8.62E-22). Next, we focused on laboratory results for N=85,460 individuals of European and African ancestry using the Vanderbilt Biobank, BioVU For the 323 laboratory traits tested, we found 32 labs significantly associated with four genes (ABO, IFNAR2, KEAP1, and SLC6A20; p < 1.55E-04). Of these, ABO captured 27/32 significant associations (mean OR = 1.33; p = 8.02E-14 < 1.40E-04). Across the 12 broad lab definitions in our data, our significant LabWAS findings were enriched for blood-related lab measurements (4.21-fold enrichment, p = 1.23E-11). Therefore, our work shows that genes implicated in COVID hospitalizations are associated coagulation-related clinical symptoms and blood-cell-related biomarkers at baseline.
Predictive Model for Inpatient Mortality

Joshua C. Smith, PhD, Allison B. McCoy, PhD, John A. Morris, Jr., MD, Ashi O. Weitkamp, PhD
Vanderbilt University Medical Center, Nashville, TN

Introduction
Over 700,000 Americans die in hospitals annually. The ability to forecast inpatient mortality soon after admission, particularly when death is not anticipated, could identify those who may benefit from additional clinical vigilance or consultation for advanced directives. A comprehensive mortality model must incorporate information from multiple domains: demographics, comorbidities, pharmacy, laboratory, and social determinants of health (SDOH). Previous models often focus on a single domain, are calculated at discharge, or only rarely include SDOH. Utilizing Electronic Health Record (EHR) data commonly available in the first 24-48 hours after admission, we are developing such a model. If successful, this project could reduce inpatient mortality, length of stay, hospital complications, and cost.

Methods
We developed a predictive model using random forests with both simple and engineered features as input. The simple features, those extracted directly from our data sources, included demographic and insurance information, counts of prior admissions and active medications, Braden score, mean arterial pressure, and common laboratory results, among others. The engineered features included Area Deprivation Index (ADI) and two popular comorbidity indices. The ADI is a measure ranking neighborhoods by socioeconomic disadvantage which has been linked with health outcomes. We calculated ADI using patients’ ZIP codes. The Elixhauser and Charlson Comorbidity Indices, morbidity/mortality predictors based on coded problems and past diagnoses, were calculated using the R package. The target output of our model was inpatient death. Our training and test sets consisted of approximately 84,000 and 44,000 historical patient admissions, respectively. As all features were not available for all patients, we developed a three-tiered system by creating slightly different models depending on available data, the major difference being the availability of certain lab results. We built our random forest models using the R package randomForest; individual models were created for each tier using a snapshot of patient data from 24-48 hours after each admission.

Results
Training and test sets contained 1832 and 1125 inpatient deaths, respectively. The AUC for tiers 1, 2, and 3 were 0.891, 0.895, and 0.893, respectively. In Table 1, we report sensitivity, specificity, and other metrics at a variety of decision thresholds. At the 0.10 threshold, for example, the model would have correctly flagged 913 inpatient deaths (81%) shortly after admission, representing 85% of expected and 80% of unexpected deaths (based on hospital mortality committee reviews) over the one-year period represented by the test set. The average time between prediction and death was just over eight days. We also analyzed retrospective alert performance in the one-year test set, as well as which features contribute most to the models.

Discussion
The model is potentially clinically impactful because it forecasts a high percentage of deaths soon after admission, but further development is required to address low precision to reduce false positives and prevent alert fatigue. Also, our use of 5-digit ZIP codes attenuated the impact of ADI for SDOH. Our next steps include calculating exact ADI from the full patient address, as well as experimenting with additional features and machine learning techniques.

References
Toward Privacy Implications in Developing and Deploying Chatbots in Informatics Research: Analysis of Chatbot Development Platforms

Diva Smriti, M.S.1, Rahil Rathod, B. Tech.1, Jina Huh-Yoo, MHCI, Ph.D.1
1Drexel University, Philadelphia, PA, USA

Introduction
Chatbots, a software application that allows for chat conversation via text or text-to-speech, are used for many purposes including information access, education, therapy, and customer service. Increasing number of researchers use chatbots for human subjects research. Research has showed consumers’ privacy concerns with chatbots1. This is because users interact with chatbots in a natural conversation form, and sensitive information can be unintentionally shared by the users as they converse with the chatbot. More work is required to examine how end-users or human subjects should be informed about how their data may be used in a complex integration of multiple third-party applications (TPAs) in enabling a chatbot. TPAs are those applications that are owned and operated by organizations external to the parent organization of the chatbot platform, that may collect end-users’ information through chatbots.

In this study, we investigated TPAs’ involvement in chatbots, and what information is presented to developers about end-user information being collected by TPAs to bring implications for human subject research with chatbots. We captured the process of creating chatbots from nine chatbot development platforms. We then conducted thematic analysis on chatbot developers’ access and control over end-user information and TPAs’ involvements.

Methods
We searched developer forums and search engines to retrieve a list of popular chatbot development platforms. From a total of thirteen platforms, we excluded the platforms that did not offer graphical user interfaces to focus on those that researchers could easily create chatbots with without in-depth programming knowledge. To standardize data collection, we also excluded voice-based only chatbots that did not offer a text-based and those that were only available to organizations not individual developers. This process resulted in nine chatbot platforms analyzed in this study: Google Dialogflow (P1), Amazon Lex (P2), ManyChat (P3), Chatfuel (P4), Wit.ai (P5), SAP Conversational AI (P6), IBM Watson (P7), Microsoft Bot Framework (P8), and QnA Maker (P9). We simulated a researcher navigating through the process of creating the chatbots, taking screenshots to document the process. We conducted inductive analysis of the screenshots in Nvivo 12 to generate themes on TPA’s involvement in chatbot development and use.

Results
Out of the 9 platforms, P3, P4, P5 required end-users to sign-in through a TPA (Facebook), but only P3, P4 explicitly state that the TPA collects end-users’ profile picture, language, and gender. Only P1 and P2 explicitly gave an option to the developers in enabling or disabling logging of end-user data. Only two platforms, P6 and P7, required developers to agree with their privacy policy, and five chatbot development platforms (P3, P4, P5, P6, P7) required developers to agree with their Terms of Service before moving ahead in the process of setting up a chatbot. These privacy policies and terms of conditions may include terms of TPAs for end-user data, the details of which were not mentioned in the set-up process of the chatbots. All nine chatbot development platforms had options for the developed chatbots to be deployed and/or exported on multiple TPAs to reach a wider audience. Developers had no control or given explicit explanation over the end-user information types being collected as part of chatbot deployment and export.

Conclusion
While for the developers the deployed platforms (e.g., Facebook) of chatbots are TPAs, for end-users, TPAs may be the chatbot development platform itself. Along with our findings, this shows the complexity of multiple TPAs being involved in chatbot development and use. The development platforms fail to explicitly state the information types collected from the end-users by TPAs. This shortcoming brings risks for conducting human subject research with chatbots, where the lack of control over human subject data pose challenges in applying key research ethics principles of data protection and privacy, such as informed consent, voluntariness, and coercion. Future work should examine how multi-parties are involved in end-users’ chatbot use and investigate end-user information types in terms of use and privacy statements across the multi-party applications. Researchers should be mindful of developing and using chatbots in human subjects research and ensure to inform IRB and human subjects on their pitfalls of end-user privacy.

References
Longitudinal Data of Cancer Patients with Prior Mental Health Diagnoses Show Differences in Demographics, Emergency Visits, and Suicidality Rates

Saira Somnay, William Chen, MD, Lauren Boreta, MD, Steve Braunstein, MD, PhD, Julian C. Hong, MD, MS

1University of California, San Francisco, San Francisco, CA, USA
2University of California, Berkeley, Berkeley, CA, USA

Purpose: Patients with pre-existing mental health conditions may be at greater risk of mental/emotional distress and need more frequent medical attention after a cancer diagnosis. The aim of this study was to compare demographics and disease characteristics/outcomes between cancer patients with pre-existing mental health conditions to those without. Further, longitudinal experience recording in a healthcare system may enable informatics-based approaches to identify patients with cancer, mental health diagnoses, and the outcomes of these comorbid conditions.

Methods: This was an electronic health record-based cohort study, utilizing an institutional Clinical Data Warehouse. A retrospective cohort study was conducted of patients with a new onset diagnosis of malignancy (ICD-10 codes C00-C97, with conversion of ICD-9 codes using the 2018 CMS general equivalence mappings) identified using an institutional de-identified electronic health data warehouse. Data collected included demographics, Charlson comorbidity index excluding cancer, and mental and behavioral health diagnosis (MHD, ICD-10 codes F00-F99), excluding substance use disorders, and emergency visits. ICD codes associated with episodes of suicidality were also extracted. Logistic and Cox regression was performed. Unless specified, models were adjusted for age, sex, cancer type, marital status, race/ethnicity, and Charlson comorbidity score.

Results: A total of 58,066 (46.6% female) patients (aged 18-90 years old) with 107,540 person-years of follow up were identified with a new diagnosis of cancer from January 2012. Of these patients, 2,683 (4.62%) had a prior MHD recorded. The most common prior MHD were depression (N= 939, 35.0%), stress related generalized anxiety (N=860, 32.1%), and Reactive and adjustment disorder (N=331, 12.3%). Based on logistic regression, patients with prior MHD were more likely to be female (OR 1.78, 95% confidence interval [CI] 1.64-1.92), have a higher Charlson comorbidity index (OR 1.61 per point, 1.58-1.65), to be unmarried (OR 1.95, 1.78-2.13) or widowed (OR 2.07, 1.87-2.40), to be a former smoker (OR 1.25, 1.15-1.35) or current heavy smoker (OR 1.42, 1.19-1.70), and were less likely to be Hispanic (OR 0.72, 0.61-0.81). After adjusting for cancer type and other covariates with Cox regression, patients with prior MHD experienced emergency visits at a slightly higher rate (HR 1.26, 1.16-1.36, P=2x10^-4). Patients with prior MHD were also substantially more likely to experience suicidality following a cancer diagnosis (N=32 of 2,683 [1.2%], OR 9.83, 5.82-16.36); overall, N=79 patients received ICD codes associated with suicidality after a cancer diagnosis, at a median of 210 days post-diagnosis (interquartile range 91.8-736.5 days). Patients receiving ICD codes associated with suicidality were less likely to be female (OR 0.58, 0.33-0.98), were younger (OR 0.97 per year, 0.95-0.99), and were significantly more likely to be unmarried (OR 3.75, 2.12-6.88) or divorced (OR 4.31, 1.94-9.17).

Conclusions: Patients with a mental health diagnosis preceding a cancer diagnosis showed increased emergency room visit rates and suicidality rates. The risk factors identified here may help identify patients needing increased medical support following a cancer diagnosis.

References:
Natural Language Processing Algorithm to Detect Terms Representing Risk of Hospitalization or Emergency Department Visits during Home Health Care

Jiyoun Song1, PhD, RN; Marietta Ojo2, MPH; Margaret V McDonald2, MSW; Kenrick Cato1, PhD, RN; Sarah Collins Rossetti1,3, PhD, RN; Yolanda Barron2, MS; Sridevi Sridharan4, MSc; Sena Chae4, PhD, RN; Mollie Hobensack1, BS, RN; Kathryn H. Bowles2,5, PhD, RN; Maxim Topaz1,2, PhD, RN

1School of Nursing, Columbia University, New York, NY; 2Center for Home Care Policy & Research, Visiting Nurse Service of New York, New York, NY; 3Department of Biomedical Informatics, Columbia University, New York, NY; 4College of Nursing, University of Iowa, Iowa City, IA; 5School of Nursing, University of Pennsylvania, Philadelphia, PA

Introduction: More than 3.4 million adults receive home health care (HHC) every year in the United States, but about one in three HHC patients experience unplanned hospitalizations or emergency department (ED) visits during HHC service.1 One strategy for risk reduction is early identification of patients at risk to prevent negative clinical outcomes (e.g., deterioration, rehospitalizations).2 Given a significant portion of risk factors are documented in narrative notes,3 natural language processing (NLP) algorithms can be used for early identification. Thus, this study aimed to test an NLP algorithm to identify terms that represent risk factors associated with unplanned hospitalizations or ED visits documented in HHC clinical notes.

Method: In previous work,4 a list of potential risk factors for unplanned hospitalizations or ED visits were identified using the problems and their associated signs/symptoms from the Omaha System4 (a standardized interdisciplinary terminology). Specifically, 32 problems including 163 individual signs/symptoms were identified: e.g., problem “health care supervision” included signs/symptoms such as “fails to obtain routine/preventive care” or “inconsistent source of health care,” among others. This list of text terms was expanded by mapping to the Unified Medical Language System (UMLS) synonyms. Then, to identify synonyms for the risk terms present in HHC clinical notes, word-embedding model (Word2Vec) was built from a large body of HHC clinical notes that were approximately 1 million notes extracted from the largest HHC agency in the Northeast United States between 2015 and 2017. For validation purpose, gold standard testing set of 500 randomly extracted clinical notes was annotated by three clinical HHC experts for presence of the 32 problems. The observed interrater agreement was 0.89 and all discrepancies were resolved through discussion. Lastly, the NLP algorithm was tested on the gold standard testing set and precision, recall, and F-score for each problem category were calculated. To prevent the error caused by the small sample size, the performance was calculated only for the Omaha System problems documented 10 times or more.

Results: In the testing dataset, 298/500 notes (59.6%) included at least one Omaha System problem (total n = 327 problems). The most common problem documented in the clinical notes was “neuro-musculoskeletal function” (13%) which included signs/symptoms of “decreased balance” and “gait/ambulation disturbance,” “skin” and “mental health” problems followed as the next most common problems documented (8%). The overall NLP system’s performance of problem identification was good (average F-score =0.84), with best results for the “social contact” problem (F-score=1) and the poorest results for the “digestion-hydration” problem (F-score=0.69).

Discussion and Conclusion: This study is the first to test an NLP system to detect terms representing risk of hospitalization or ED visits in HHC. The NLP system achieved good performance on a subset of randomly selected HHC clinical notes. Our findings suggest that information regarding the risk of hospitalization or ED visits was frequently documented in narrative clinical notes and NLP can extract information from narrative clinical notes to improve our understanding of care needs in HHC. Further, these results can inform a development of an early warning scoring system to help reduce negative outcomes in HHC.

Acknowledgments: This study was funded by AHRQ #R01HS027742: Building risk models for preventable hospitalizations and emergency department visits in homecare (PI: Maxim Topaz).

References

Artificial intelligence identifies the progression of cancer patients with kidney disease

Qianqian Song¹,², Jing Su³
¹Center for Cancer Genomics and Precision Oncology, Wake Forest Baptist Comprehensive Cancer Center, Wake Forest Baptist Medical Center; ²Department of Cancer Biology, Wake Forest School of Medicine, Winston-Salem, NC, USA; ³Department of Biostatistics and Health Data Science, Indiana University School of Medicine, Indianapolis, IN, USA

Introduction As many cancer chemotherapeutic agents causes kidney dysfunction, acute and chronic kidney disease often manifests during different cancer stages. Given the potential link between kidney disease and cancer, investigating the landscape of the progression trajectories in the Cancer Patients with Kidney Disease (CPKD) allows us to better understand their specific progression patterns and heterogeneity among this risk population, as well as the underlying associated risk factors. Electronic medical records (EMRs) covering large CPKD cohorts with demographic, clinical, and socioeconomic features, provide unique and valuable opportunities to capture patients’ developmental status [1]. Through utilizing the graph-based artificial intelligence and a novel graph learning algorithm on EMRs data, we aim to learn the progression trajectory and the risk factors, which will contribute to the therapeutic improvement of patients with both cancer and renal impairment.

Methods and Results In this study, we collect the electronic medical records (EMRs) data at Wake Forest Baptist Medical Center (WFBMC), with 79,434 patients and 508,733 clinical encounters, to outline the progression trajectory roadmap of CPKD. Patients’ health status is described by 18 Essential Clinical Indices across clinical encounters. As the EMRs data are heterogeneous in comorbidities and clinical conditions, we first learn the deep latent representation of the heterogenous raw data through the graph-based artificial intelligence model, i.e., GraphSAGE [2](Figure 1A). Based on the deep represented EMR data, we then implement a novel graph learning algorithm, Discriminative Dimensionality Reduction Tree (DDRTree), to establish the progression trajectories of CPKD (Figure 1B). We also verify that this progression trajectory can be accurately predicted using the k-nearest neighbor model. Importantly, with the learned graph structure, we then annotate the trajectories with clinical, genomic features, and major risk indices of cancer and kidney function. Patients at the high-risk trajectory are associated with known modifiable risk factors such as hypertension, suggesting that the learned trajectories are clinically relevant. Meanwhile, novel and actionable knowledge about preventing the transition from low-risk to high-risk conditions are learned, including drugs with previously unknown nephrotoxicity.

Conclusion The progression trajectory roadmap reveals and highlights the disease progression patterns of CPKD patients. We identify the risk factors and prognostic clinical indicators associated with discovered trajectories. The CPKD progression trajectory roadmap reveals diverse kidney failure pathways associated with different clinical conditions. Specifically, we identify one high-risk trajectory and one low-risk trajectory. Patients who are associated with the high-risk trajectory demonstrate fundamentally different disease progression mechanisms from those at the low-risk trajectory.

References
Deploying Conversational Agents to Facilitate Housing Assistance Needs Resulting from COVID-19

Brett R. South MS PhD1, Anita M. Preininger PhD1, Piyush Parmar1, Rubina Rizvi MD, PhD1, David Brotman MS1, Shira Alevy EdM1, Mollie McKillop MPH PhD1, Gretchen Purcell Jackson MD PhD1,2, William Kassler MD1

1IBM Watson Health, Cambridge, MA, USA; 2Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN USA

Introduction
Nevada’s CARES Housing Assistance Program (CHAP) was implemented in Clark County, Nevada to address employment and financial challenges faced by residents as a result of the COVID-19 pandemic. To enhance access to resources and benefits, Clark County deployed a conversational agent, Watson Assistant (WA), to answer residents’ questions about the program, reduce web traffic to the application website, and screen for program eligibility.1 The conversational agent directs residents who may qualify for assistance to a website where citizens can apply for program benefits. Qualified beneficiaries include people residing in Clark County who have evidence of a pending eviction and who have less than $3,000 in a savings or checking account, with an annual household income below $118,000 or 120% of the median area income in the 2020 CARES Act program and below $79,200 or 80% in the 2021 Emergency Rent Assistance program.

Methods
We conducted a descriptive analysis of user interactions with the CHAP conversational agent and assessed utilization in the first three months of use (between 10/2020 and 12/2020). We also quantified the number of messages and conversations over the most recent 30-day time period. The WA platform can be used to build a conversational interface into a website or automated voice system. The platform can be trained to include customized information related to any specific use case, language preference, location or organizational content. There are four natural language capabilities of WA: (1) understanding content, (2) topic classification, (3) retrieving information from a knowledge base, and (4) generating responses in natural language. When a user enters questions about housing assistance, the conversational agent interprets the question to identify the intent (target of a user’s query) and match it to an internal list of topics that best answer the user’s question. This conversational agent is currently trained to support 29 core topics derived from policy documents including eviction, eligibility, benefit amounts, application status, application approvals, and appeals.

Results
The conversational agent created for the CHAP program was used by over 66,000 users to answer common questions regarding housing assistance in Clark County, Nevada over the three-month study period. In the month between December 13, 2020 and January 12, 2021, the system handled over 230,901 conversations (an average of 1.46 messages per conversation), and 348,521 messages. The most frequent topics of users’ queries over the study period included eligibility status (21,555), coverage (6,766), CARES program summary (5,241), and documentation (1,692).

Conclusion
This study describes successful deployment of a conversational agent designed to assist residents in navigating housing assistance programs in Clark County, Nevada during a public health emergency, demonstrating that WA can be customized to fulfill a variety of informational needs. Conversational agents have the potential to aid in screening a large number of applicants for program eligibility during periods of high need. The ability to answer user questions efficiently may facilitate enhanced delivery of information to a large number of individuals and may increase users’ access to benefits and services.

References
Automatic Stratification of Tabular Health Data

Skyler Speakman, PhD¹,* , Girmaw Tadesse, PhD¹, Victor Akinwande¹, William Ogallo, PhD¹, Claire Mershon, MPH², Nosa Orobaton, MD, PhD², Daniel B. Neill, PhD³

¹IBM Research, Nairobi, Kenya; ²Bill & Melinda Gates Foundation, Seattle, WA; ³New York University, NY

Stratifying an outcome of interest across a categorical feature is one of the most commonly-used methods of data visualization and exploration. This is because analyzing data through sub-groups is immediately intuitive to a broad range of audiences. Stratification has low barrier-to-entry data requirements and is applicable to almost any form of tabular data. The goal of stratification is to identify sub-groups where the outcome of interest (e.g., mortality, disease burden, access to care) is significantly higher or lower than the overall global average. Identifying these sub-groups is a key first step toward better understanding of complex data and formulating initial hypotheses as to the relevant trends and patterns. AutoStrat (Automatic Stratification) builds on the universal strengths and simplicity of manual stratification while adding high-dimensional scalability and disciplined, data-driven discovery. Manual stratification requires the investigator to presuppose the categorical feature to view the outcome of interest through. This may be repeated for multiple features separately in a manual process. In some cases, two features (e.g., age decile and gender) may be used together. However, going beyond two or three features introduces computational complexity from the exponentially-large search space of possible sub-groups to stratify across. AutoStrat efficiently stratifies over multiple features simultaneously and returns the strata (i.e. sub-group or sub-population) that has unexpectedly high (or low) outcomes as compared to the mean. AutoStrat builds on the subset scanning literature and Bias Scan in particular. We frame AutoStrat as a search problem with the goal of finding the “most anomalous” subset of the data by identifying which sub-population a naive predictive model is most biased against.

AutoStrat was applied to the Alliance for Maternal and Newborn Health Improvement (AMANHI) study data and focus on ~21,000 women/births in Ghana. 1.53% of these births resulted in a neonatal mortality. The identified strata was births at home where either a doctor or midwife was the delivery person. This group included 361 births with 153 deaths (42.4%). This group had a 28-fold increase in the proportion of neonatal mortality as compared to the overall proportion of 1.53%. This sub-population contained 46% of all neonatal deaths in the Ghana study. AutoStrat differs from explaining predictions made by machine learning models that have learned feature interactions. Rather, it discovers which sub-population shows the most evidence of higher outcomes than the global average without any predictive modeling assumptions in place. This pivot from fitting regression models (or interpreting black-box machine learning models) to understanding data through anomalous sub-populations represents an important direction in medical informatics and data science research. Code for AutoStrat is publicly available through the AI Fairness Toolkit. This work is funded by Bill & Melinda Gates Foundation, investment ID 52720.

References

Amanda Steffon, LCSW, Precious Faith Porter, RN, BSN
3M Health Information Systems, Murray, UT, USA

Introduction
The World Health Organization (WHO) defines health as “a state of complete physical, mental, and social well-being and not merely the absence of disease or infirmity.” Social Determinants of Health (SDoH) aim to address the external factors that impact health outcomes. As healthcare moves to address SDoH, data from Social Work (SW) should be incorporated. Social workers have been addressing disparities in our communities for over a century. They collect extensive and comprehensive information through assessment and evaluation during the provision of care. The SW record captures accurate and pertinent information that could contribute data to SDoH. This study aims to identify what information is currently available in SW records that can be collected and aggregated for SDoH.

Methods
SDoH categories and concepts were identified by subject matter experts (SMEs). To assess the contribution of SW documentation to SDoH information, we reviewed a sample set of SW psychosocial assessments that were available to the lead author and encoded the extracted SDoH concepts using SNOMED CT. Web-based searches were conducted to gather collateral information and validate team discussions. Findings were documented and reviewed by SMEs to reach consensus.

Results/Discussion
The following sections in a standard SW psychosocial assessment were identified: Demographics, Education, Employment, Family Support, Social History, Medical Conditions and Precipitating Events, Discharge Planning, Mental Health/Substance Abuse, Hospitalizations, Recommendations/Referrals, and Patient Strengths. Though these topics are occasionally mentioned in provider documentation, they are consistently documented in SW assessments. In the following table, we provided two examples encoded to SNOMED CT.

<table>
<thead>
<tr>
<th>SW section</th>
<th>Language in psychosocial assessment</th>
<th>Coding (SNOMED CT)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Condition precipitating admission</td>
<td>Pt was found in front of a 7-11 having uncontrollable seizures. The shelter clinic reported he had not filled his seizure medication for over a month. Pt reported that he was unable to afford his refill.</td>
<td>715036000</td>
</tr>
<tr>
<td>Discharge planning</td>
<td>Pt is chronically homeless and reports he does not have any family. He only receives $450 a month in Social Security Disability. SW also got pt on the waiting list for a Housing Choice Voucher (HCV).</td>
<td>329110000</td>
</tr>
</tbody>
</table>

On average, a SW assessment resulted in 12-20 unique SDoH concepts. Unfortunately, SW assessments are inconsistently formatted and stored in EHRs: Free text notes, MS Word documents, faxes, surveys, screening tools, etc. Even when an EHR SW assessment module is available, the data collected is most often not encoded, let alone standardized or interoperable. While EHR vendors are more frequently incorporating SW assessments into their platform and involving SW in the development, there is no standard for data collection and extraction. Industry efforts to develop a consensus list and organizational efforts to create terminology are a start to ensure that the necessary SDoH concepts are created. We need to look forward to how SDoH data will be captured from patients and how this aligns with SW practice. While the development effort is not trivial, it may still be more realistic than placing the burden of collecting SDoH on already overworked physicians. Further studies could be conducted on extracting and encoding SW assessments for SDoH compared to the data encoded from providers in EHRs.

Conclusions
Healthcare organizations that utilize social workers are frequently in possession of large amounts of SDoH data but are not extracting, analyzing, or drawing meaningful conclusions from it. There continues to be a need for alignment between the clinical workflow of social workers and EHR workflow while closing the gaps between social services and providers. It is critical to involve social workers in the development and design of standard terminologies and software systems while advocating for the SDoH movement.
Comparing Language Model Vocabulary Coverage on Clinical Documents

Bryan D. Steitz, PhD; Adam Wright, PhD
Dept. of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN

Introduction
Large bidirectional language models have transformed the field of natural language processing (NLP). These models utilize tokenized word sequences as inputs, and context-dependent word embeddings, which enables them to capture semantic meanings in text. The degree to which the language models can calculate accurate word embeddings depends on the extent to which their internal vocabulary matches text to be processed. A key example of a large bidirectional language model is BERT, which was trained on Wikipedia and a set of over 11,000 unpublished books. Many published models have retrained word embeddings on a domain-specific corpus (for example, biomedical text), but most of them have used the base BERT vocabulary, leading to many unknown input words. The goal of this study was to train a new vocabulary on a corpus of discharge summaries and compare coverage with popular language model vocabularies on corpora of clinical documents.

Methods
To develop our vocabulary, we extracted a random sample of 150,000 discharge summaries from January 2019 to February 2021 from Vanderbilt University Medical Center’s Epic electronic health record (EHR) system. We preprocessed the text to remove accent characters and split the text on whitespaces and punctuation. We trained cased and uncased 30,522-item wordpiece vocabulary with the preprocessed text using the HuggingFace Transformers library.1 We compared our vocabulary with the cased and uncased SciBERT2 and BERT Base3 vocabularies on random samples of 10,000 discharge summaries, progress notes, and patient instructions. We preprocessed each note by removing non alphanumeric characters. To test uncased vocabularies, we also converted all words to lowercase. For the purposes of comparing coverage, we tokenized each clinical document in its entirety, regardless of length.

Results
The average document length was 895.0, 527.9, and 493.3 tokens for discharge summaries, progress notes, and patient instructions, respectively. We present in Table 1 the percentage of unknown tokens encoded by each vocabulary. BERT and SciBERT had much lower success in matching clinical documents compared to our domain-specific vocabulary.

Table 1: Percentage of Unknown Tokens by Vocabulary and Document Type

<table>
<thead>
<tr>
<th></th>
<th>Discharge Summaries</th>
<th>Progress Notes</th>
<th>Patient Instructions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Uncased Vocabularies</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BERT Base</td>
<td>16.1%</td>
<td>15.3%</td>
<td>9.3%</td>
</tr>
<tr>
<td>SciBERT</td>
<td>12.1%</td>
<td>11.7%</td>
<td>9.8%</td>
</tr>
<tr>
<td>VUMC DischargeSumm</td>
<td>0.0%</td>
<td>0.0%</td>
<td>1.4%</td>
</tr>
<tr>
<td><strong>Cased Vocabularies</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BERT Base</td>
<td>28.4%</td>
<td>26.4%</td>
<td>14.8%</td>
</tr>
<tr>
<td>SciBERT</td>
<td>1.0%</td>
<td>2.6%</td>
<td>2.8%</td>
</tr>
<tr>
<td>VUMC DischargeSumm</td>
<td>0.0%</td>
<td>0.0%</td>
<td>1.2%</td>
</tr>
</tbody>
</table>

Conclusion
Adequate vocabulary coverage is integral to obtain accurate contextual word embeddings. Previous work has found that applying a domain-specific vocabularies can improve NLP performance across multiple tasks.2 We hypothesize that training domain-specific and site-specific vocabularies and word embeddings can further improve state-of-the-art performance and transfer learning, which we will test in future work.

References
Use of the Electronic Health Record in Pediatric Medicine to Capture and Highlight Gender Identity in Order to Better Ensure Gender Affirming Care

Stacey T. Stokes, MD MPH1, Jessica Herstek, MD1

1Children’s National Hospital, Washington, DC

The number of gender diverse youth is growing in the United States, and with higher rates of negative health outcomes including depression and suicidality, at-risk behaviors, and bullying compared to their cis-gender peers, this cohort is a two-fold vulnerable population that is currently not having their needs met in the pediatric healthcare community. While documentation of gender identity (GI) is explicitly targeted within federal and academy initiatives, pediatric institutions and providers have been slow to respond. Many transgender patients are still educating their providers on GI language, and most pediatricians do not ask GI-specific questions during visits. This leads to a patient-provider relationship that extends the non-affirming environment and has impacts on patient well-being and interactions with healthcare professionals. Institutions can take several steps to ensure equity when considering gender affirming care (Table 1).

Table 1: Steps for Healthcare Institutions to Consider in Providing Gender Affirming Care

<table>
<thead>
<tr>
<th>Step</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Buy-in from stakeholders and leadership to support gender affirming language adoption</td>
</tr>
<tr>
<td>2.</td>
<td>Implement a document component for gender identity (GI) that allows for face-up visibility within the electronic health record (EHR): Capture affirmed name, gender identity, birth sex, and preferred pronoun at minimum.</td>
</tr>
<tr>
<td>3.</td>
<td>Standardize documentation capture of gender identity components</td>
</tr>
<tr>
<td>4.</td>
<td>Educate all staff on gender affirming language to encourage empathic GI conversations</td>
</tr>
<tr>
<td>5.</td>
<td>Obtain feedback from cis- and trans-gender patients and families on whether needs are being met. Adapt as needed.</td>
</tr>
</tbody>
</table>

Gender Identity Documentation Implementation: A Proactive Model

The objective of our project was to uplift the universal demographic banner bar in our EHR to highlight GI-related content and allow providers and nursing staff the ability to quickly edit the information for real-time updates. Our clinical informatics team collaborated with our organization’s Diversity, Equity and Inclusion (DEI) taskforce and our EHR vendor (Cerner, Kansas City) to integrate both custom and standardized tools to capture sexual orientation, gender identity, affirmed name, and preferred pronouns, in addition to birth sex and assigned name, and have the information display face-up on the banner bar for clinical and administrative staff. The unique partnership in the spring of 2021 involving three stakeholder groups allowed for language integration and understanding that would not have been possible otherwise. Physician informaticists designed the banner bar layout, DEI subject matter experts (SMEs) reviewed content and education language, and EHR architects implemented the changes. Buy-in from DEI SMEs and hospital leadership was essential to this change success. Further refinement of the functionality and staff education are ongoing to ensure gender-affirming language is used and captured in an empathic and inclusive way to create a positive patient-provider relationship.

The data not only allows for better care which has several patient-centered benefits, but aims to improve individual and population health outcomes, aid in LGBTQ+ research, and improve information sharing for public health interventions to reduce disparities. Pediatric health care providers should prioritize GI understanding and documentation to ensure gender-affirming care, and therefore overall care, for all patients.

References
Infographic use leads to better health outcomes among Latinos with HIV

Samantha Stonbraker PhD, MPH, RN1, Gabriella Sanabria MEd2, Maureen George PhD, RN, AE-C3, Silvia Amesty MD, MPH, MSEc4, Ana F. Abraído-Lanza PhD5, Peter Gordon MD6, Susan Olender MD6, Tawanda Rowell-Cunsolo PhD7, Sophia Centi MPH1, Bryan McNair MS, PStat®1, Suzanne Bakken PhD, RN3, Rebecca Schnall PhD, MPH, RN-BC3

1University of Colorado College of Nursing, Anschutz Medical Campus, Aurora, CO; 2University of South Florida College of Public Health, Tampa, FL; 3School of Nursing, 4Irving Medical Center, Columbia University, New York, NY; 5College of Global Public Health, New York University, New York, NY; 6New York Presbyterian Hospital, New York, NY; 7University of Wisconsin – Madison School of Social Work, Madison, WI

Introduction

The effective visualization of information is a key informatics process, yet methods to use visualizations for clinical communication is understudied. Well-designed infographics (graphics that convey information) have the potential to augment communication regarding Human Immunodeficiency Virus (HIV) by providing complex information in simple, visually appealing ways. This can be useful when patients and providers have different cultures/languages, as is the case for many Latino persons living with HIV (PLWH) who receive care in the United States. This disconnect between patient and provider is just one contributor to the HIV-related disparities experienced by Latino PLWH, and may be improved through infographic use. The purpose of this study was, therefore, to assess if using infographics to convey health information with Latino PLWH can improve communication and lead to better health outcomes.

Methods

We conducted a longitudinal pre-test post-test study at a clinic in Manhattan. We trained two bilingual researchers to use infographics (English/Spanish) that contain common HIV-related concepts and were previously designed by our team.1 Then, we enrolled 30 adult Latino PLWH. After three of their normal appointments, a researcher offered health information specific to their needs using infographics. At each session, we extracted CD4 count and viral load from medical records (primary outcomes) and participants completed questionnaires containing validated scales to measure HIV-related knowledge, satisfaction with provider, self-efficacy to manage HIV, adherence, and self-reported health status (secondary outcomes). Outcomes were selected using a theoretical model that identifies ways communication may influence health outcomes.2 We first summarized all variables with descriptive statistics. Then, we conducted chi-square tests for the categorical outcomes and Kruskal-Wallis nonparametric tests with the predictors: visit number (1-3), new patient status, and if visits were pre or post the COVID-19 research pause, for continuous outcomes.

Results

Participants (N=30) were mostly male (60%) and Spanish-speaking (60%), with a mean age of 52.4 years (SD±14.2). Changes in CD4 count and viral load were not significantly different. Knowledge scores improved between visits (p<.0001 for both time points). Health status also improved over time (p=.02). Knowledge scores (p=.04) and current health status (p=.01) also changed significantly from before to after the COVID-19 research pause.

Conclusion

This study provides preliminary evidence that infographics may improve communication and health outcomes among Latino PLWH. Observed changes are likely to be clinically significant, as participants demonstrated more understanding of self-management strategies and reported better health, but this must be further studied. The initial success of this intervention and questions it uncovered warrant robust exploration in a large, randomized trial.

References

Service Delivery Adaptations: Telehealth Creative Arts Therapy in the Veterans Health Administration

Kristin M Story¹, Sheri L Robb², Alaina Preddie ¹,³, Mindy Flanagan¹, Teresa M Damush¹,³,⁵

¹Department of Veterans Affairs (VA), Richard L. Roudebush VA Medical Center; Indianapolis, IN; ²Indiana University School of Nursing; ³VA Health Services Research and Development (HSR&D) Expanding Expertise through E-health Network Development (EXTEND) Quality Enhancement Research Initiative (QUERI); Indianapolis, IN; ⁴Department of Medicine, Indiana University School of Medicine; ⁵Regenstrief Institute, Inc

Introduction: In response to the COVID-19 pandemic, the Veterans Health Administration (VHA) instituted a Nation-wide shift to increase virtual care.¹,² However, there are unique technical challenges to telehealth delivery of Creative Arts Therapies (CAT) that include music and art. The aims of this study were to identify how telehealth CAT is being employed, clinician level barriers and facilitators, and how clinicians have adapted their practice for telehealth CAT delivery.

Methods: We deployed a survey, guided by the Consolidated Framework for Implementation Research (CFIR), to all VHA facilities.³ Of the 120 CAT therapists surveyed, 92 (77%) responded and a subset of 23 therapists were interviewed. Quantitative data from the surveys were analyzed. Open field responses were summarized and coded for thematic analysis. Interviews were video recorded and transcribed. De-identified transcripts were used for coding and thematic analysis by a minimum of two coders from the research team.

Results: Survey findings indicate that 76% of survey respondents have delivered a telehealth CAT session with 74% delivering more than 50 in the last year in response to the pandemic. More than 85% of CAT therapists have adapted or created new interventions in order to deliver services through telehealth. Most therapists (65%) encountered moderate technology issues that they were able to overcome and 55% of therapists indicated that technology issues were specific to their CAT discipline. Open field analysis from the surveys revealed challenges related to sound quality (for music therapists) and visual limitations (for art therapists). Administrative requirements to use specific platforms and software increased the need for CAT work arounds to deliver sessions successfully. But CAT therapists also identified increased creativity and clinician growth from telehealth delivery as well as the benefits of interacting virtually in the Veteran’s home environment. Organizational facilitators included new interventions/directives, and challenges related to specific platforms (e.g., sound/visual range).

Discussion – Identification of key facilitators and barriers to implementation and delivery has allowed us to create a foundational tool kit comprised of information and resources to help CAT therapists in their delivery of services. The relative advantage of CAT virtual delivery to reach patients across VHA is now being adopted by the healthcare organization as a hybrid delivery approach to improve its reach into its patient population.

References

A Machine Learning Approach for Identifying Emergent Phenotypes Associated with a Previous COVID Infection

Zachary H. Strasser, MD, MBI1,2, Hossein Estiri PhD1,2, Shawn Murphy MD, PhD1,2,3
1Massachusetts General Hospital, Boston, MA, 2Harvard Medical School, Boston, MA, 3Mass General Brigham, Boston, MA,

Introduction
The SARS-CoV-2 virus has long-lasting effects on a variety of organ systems.1,2 However many of the large cohort studies used for quantifying the symptoms and diseases that follow an acute COVID-19 infection, rely on large insurance databases or electronic health record systems that are at risk for inaccuracies, missingness, and incompleteness. Accurate identification of phenotypes will help the healthcare system focus its efforts and resources on adequately controlled age- and gender-specific sequelae of a COVID-19 infection.

Methods
We applied a computational framework for knowledge discovery from clinical data, called MLHO3 (Machine Learning for predicting Health Outcomes), to identify phenotypes that: 1) appeared for the first time in a patient’s electronic health records (EHRs) after the acute phase of the infection and 2) are positively associated with a past positive reverse transcription-polymerase chain reaction (RT-PCR) test for COVID-19. We evaluated the post-test phenotypes (based on International Classification of Disease codes) in two temporal windows after the test and in subpopulations based on age and gender. We then validated these new phenotypes with a physician’s chart review of the clinical notes to ensure that the diagnostic labels match the real-world experience of the patient.

Results
We analyzed 96,025 non-hospitalized patients who had a COVID-19 RT-PCR test, of which 22,475 (23.41%) had a positive COVID-19 result. We identified 33 phenotypes among different age/gender cohorts or time windows that were positively associated with past SARS-CoV-2 infection. All identified phenotypes were newly recorded in patients' medical records two months or longer after a COVID-19 RT-PCR test in non-hospitalized patients regardless of the test result. We then performed a chart review of samples of the selected phenotypes to ensure the clinical notes corresponded with the emergence of a new phenotype. Among these phenotypes, a new diagnosis record for anosmia and dysgeusia (OR: 2.60, 95% CI [1.94 - 3.46]), alopecia (OR: 3.09, 95% CI [2.53 - 3.76]), chest pain (OR: 1.27, 95% CI [1.09 - 1.48]), chronic fatigue syndrome (OR 2.60, 95% CI [1.22-2.10]), shortness of breath (OR 1.41, 95% CI [1.22 - 1.64]), pneumonia (OR 1.66, 95% CI [1.28 - 2.16]), and type 2 diabetes mellitus (OR 1.41, 95% CI [1.22 - 1.64]) were some of the most significant indicators of a past COVID-19 infection. Additionally, more new phenotypes were found with increased confidence among the cohorts who were younger than 65.

Discussion
Our approach avoids a flood of false positive discoveries while offering a more robust probabilistic approach compared to the standard linear phenome-wide association study (PheWAS). More than 63 percent of PASC phenotypes were observed in patients under 65 years of age, pointing out the importance of vaccination to minimize the risk of debilitating post-acute sequelae of COVID-19 among younger adults. As part of the NIH’s RECOVER Initiative, our team at Massachusetts General Hospital serves as part of the PASC Data Resource Core (DRC) to support and contribute to the collection, coordination, and analysis of data collected on PASC patients throughout the nation facilitating analysis of standardized data across different RECOVER cohort studies. The above approach will help to inform this ongoing initiative.

References
Measuring the correctness of All of Us physical measurement

Lina Sulieman, PhD¹, Karthik Natarajan, PhD², Qingxia (Cindy) Chen PhD³, Robert J Carroll, PhD¹, Kayla Marginean¹, Paul Harris, PhD¹, Andrea Ramirez, MD, MS⁴

¹ Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, US; ² Department of Biomedical Informatics, Columbia University, New York, New York, US; ³ Department of Biostatistics, Vanderbilt University Medical Center, Nashville, TN, US; ⁴ Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, US

Introduction

The quality of electronic health records (EHR) data can affect credibility, validity, and reproducibility of research,¹ therefore assessing the quality of the research data is essential. We assessed correctness, a data quality dimension, and identified possible incorrect values for six common physical measurements in the All of Us Research Program, a multi-site ongoing initiative to recruit one million or more diverse participants.

Methods

We investigated height, weight, body mass index (BMI), heart rate (HR), systolic blood pressure (SBP), and diastolic blood pressure (DBP), extracting measures taken at study enrollment (i.e., onsite measurement) or recorded at age 18 years or older and in an outpatient visit in the EHR. To identify erroneous measurements, we used three methods. We compared the enrollment values to the corresponding EHR value within one year before or after enrollment. For 2-SD method, the sample standard deviation (SD) was calculated using onsite values from all participants, and a value outside 2*SD was labeled as erroneous. The 3-SD-participant method calculated SD for all EHR values per participant and labeled as erroneous the values that were higher or lower than 3-SD from their mean. The moving average method compared a value to the moving average of previous values. It labeled the value as erroneous if the difference between the value and moving average was higher than a threshold. We performed a grid search to select the final set of parameters for moving average. To evaluate the concordance of values between the EHR and onsite, we calculated the Pearson and Spearman correlation between median of EHR values within a year and onsite values.

Results

The number of participants in fourth release who have EHR values outside the error or 2-SD bar were 4110, 1694, 1997, 266322, 231394, and 238420 for height, weight, BMI, heart rate, SBP, and DBP. Some weight and BMI values were clustered along the lines with slopes 0.45 and 2.2, as shown in Figure 1. Height, weight, and BMI had the highest correlation. The 3-SD-participant identified around 500,000 height and 6,000 weight values in the entire EHR that might be incorrect. The moving average labeled 220,435 EHR height values. Moving average labeled 400,000 weight values as erroneous. The correlation between onsite and EHR was above 0.96 for height, weight, and BMI, while it ranged between 0.55 to 0.62 for SBP, DBP, and heart rate. After removing values labeled as possible error, the correlation for most of the values increased; however, the improvement was not significant for all measurements and it ranged between 0.002 to 0.1.

Discussion

Our study is one of the few attempts to identify erroneous measurements recorded during adulthood for six common measurements. Removing some erroneous measurement improved correlation between onsite and EHR measurements recorded within a year. Moving average and 3-SD-participants can identify incorrect measurements that are not limited by their entry dates. For heart rate and blood pressure, at least one value per participant was flagged. The intra-variability in those measurements is high since they depend on rest and stress levels, and underlying conditions. Minimizing the error rate in clinical research dataset can lead to more accurate results. Our analysis can identify mapping errors and implemented as a processing step prior training models.

References

1. Weiskopf NG, Bakken S, Hripcsak G, Weng C. A Data Quality Assessment Guideline for Electronic Health Record Data Reuse. eGEMs (Generating Evid Methods to Improve patient outcomes). 2017
Generating Longitudinal Synthetic EHR Data with Recurrent Autoencoders and Generative Adversarial Networks

Siao Sun, M.S.¹, Fusheng Wang, Ph.D.¹, Sina Rashidian, Ph.D.¹, Tahsin Kurc, Ph.D.¹, Kayley Abell-Hart, B.S.¹, Janos Hajagos, Ph.D.¹, Wei Zhu, Ph.D.¹, Mary Saltz, M.D.¹, Joel Saltz, M.D., Ph.D.¹

¹Stony Brook University, Stony Brook, New York, USA

Introduction

Synthetic electronic health records (EHR) can facilitate effective use of clinical data in software development, medical education, and medical research without the concerns of data privacy. We propose a novel Generative Adversarial Network (GAN) approach, called Longitudinal GAN (LongGAN), that can generate high-quality longitudinal clinical data containing continuous laboratory and medication values for given diseases.

Methods

We randomly chose two facilities from the 10 highest volume inpatient facilities in Cerner Health Facts database and extracted inpatient encounters from 1/1/2016 to 12/31/2017 for the experimental evaluation of the proposed method. The architecture of LongGAN is illustrated in Figure 1. A recurrent autoencoder is trained with real longitudinal data and used to extract representations, the GAN model is trained with representations of real data and used to generate synthetic representations, which are then fed to the decoder part of autoencoder to get synthetic longitudinal data. Soft labels of diseases are used as conditional input of GAN model.

Results And Conclusions

We evaluated LongGAN with Train on Synthetic and Test on Real task, where we trained two classifiers for acute kidney disease, one using the real training dataset and the other using the synthetic dataset, and then evaluated both models on a real test dataset. The classifiers we chose include logistic regression, random forest, and LSTM network. The results are shown in Table 1. Our results demonstrate that the models trained on synthetic datasets generated by LongGAN have performances closer to those trained on real datasets than other synthetic datasets generated by RCGAN and TimeGAN, two most relevant methods for longitudinal data generation.

Table 1. Performance of various predictive models on real and synthetic datasets.

<table>
<thead>
<tr>
<th>Predictive model</th>
<th>Metric</th>
<th>Real</th>
<th>RCGAN</th>
<th>TimeGAN</th>
<th>LongGAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic Regression</td>
<td>AUROC</td>
<td>0.86</td>
<td>0.57</td>
<td>0.61</td>
<td>0.71</td>
</tr>
<tr>
<td>AUROC</td>
<td>0.53</td>
<td>0.34</td>
<td>0.30</td>
<td>0.31</td>
<td></td>
</tr>
<tr>
<td>Random Forest</td>
<td>AUROC</td>
<td>0.86</td>
<td>0.50</td>
<td>0.71</td>
<td>0.72</td>
</tr>
<tr>
<td>AUROC</td>
<td>0.30</td>
<td>0.28</td>
<td>0.30</td>
<td>0.31</td>
<td></td>
</tr>
<tr>
<td>LSTM network</td>
<td>AUROC</td>
<td>0.83</td>
<td>0.43</td>
<td>0.47</td>
<td>0.72</td>
</tr>
<tr>
<td>AUROC</td>
<td>0.43</td>
<td>0.43</td>
<td>0.45</td>
<td>0.52</td>
<td></td>
</tr>
</tbody>
</table>

References

Extraction of Ambiguous Phrases Found Within Adverse Drug Event Mentions Using A Natural Language Processing-Based Annotation tool

Zhoujun Sun BS student¹, Rubina Rizvi MD, PhD², Shilo Anders, PhD¹, Brett South MS, PhD², Elisabeth L. Scheufele MD, MS², Karlis Draulis³, Henry J. Feldman MD, MS²

¹Department of Psychology, Vanderbilt University, Nashville, TN, USA; ²IBM, Watson Health, Cambridge, MA, USA

Introduction

About 80% of Electronic Medical Record (EMR) data are unstructured, rendering them relatively inaccessible to traditional analytics without manual review.¹ Natural Language Processing (NLP) offers a method to extract and classify unstructured text such as adverse drug events (ADE), drug names into structured codable concepts. ADE are defined as an injury resulting from medical intervention related to a drug.² Identifying true positive phrases with an accurate representation of a cause (drug) and effect (ADE) relations is a particularly challenging NLP problem. Ambiguity may arise from the fact that clinical notes are often written as telegraphic, short sentences, having acronyms and inconsistencies in punctuation and grammar. This could lead to uncertainty and variable interpretation, hampering communication and potentially resulting in undesirable clinical outcomes.³ We report on methods used to extract and identify a corpus of commonly observed, ambiguous phrases from clinical notes using The Watson™ Annotator Service for Clinical Data (ACD).

Methods

The MIMIC III v.1.4 database containing approximately 1.3 million full-text notes was utilized as our initial corpus. Our study focuses on ADE from admission notes for ICU patients (excluding in-hospital transfer), written by resident physicians. The MIMIC III database was loaded into MySQL, hosted on an Apple Macintosh server utilizing a 4-drive thunderbolt-3 RAID array. The notes were filtered on the NoteEvent.category for those containing “resident” and “admission”. The admission information linked to the note provided the source of the patient and excluded notes where originating service was not the emergency department. The ACD Medication Adverse Event engine processed the notes, excluding ADEs with a system probability of < 60%. Final system output included medication/agent ascribed to the ADE with surrounding text limit of +/- 100 characters and preceded by an “on” clause e.g., GI bleed on Coumadin. We removed any duplicate phrases. Using the output, we then conducted two rounds of manual, human review. In the first round, two co-authors, (ZS, RR) identified ambiguous phrases applying several heuristics (i.e., a complete, meaningful sentence with common drug mentions, removal of duplicate phrases). Next, two physicians (ELS, HF) reviewed and marked the phrases considered most concerning, and with potential critical outcomes, inter-rater agreement (IRR) was calculated. A frequency table was generated for categorical data, drug names (Table-1).

Results

The initial corpus selected from MIMIC- III database was comprised of N=11,500 admission notes from ICU admissions corresponding to an adverse drug event. The NLP algorithm identified N=9057 notes containing a possible ADE. After filtering on potential ADE phrase and applying a character limit (+/- 100 chars), the list consisted of N=597 phrases that was further reduced to the final selected N=155 after two rounds of manual review. The calculated IRR was of 89% with Kappa values (0.73 95% CI 0.8-0.87). The number of unique drug names (brand or generic) associated with an ambiguous phrases was N=102, with Coumadin having the most mentions N=112 (40%) out of 295 (Table 1).

Conclusion

In this study we demonstrate a method to identify ambiguous phrases found within ADE mentions from a collection of notes. In future efforts, these phrases will be utilized to study and model users’ (e.g., clinicians and non-clinicians) interpretation and deduction processes.

References


Table 1-Common drug mentions

<table>
<thead>
<tr>
<th>Drug name</th>
<th>N (%)</th>
<th>Categorical data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coumadin</td>
<td>112</td>
<td>(40%)</td>
</tr>
<tr>
<td>Warfarin</td>
<td>12</td>
<td>(4%)</td>
</tr>
<tr>
<td>Levitiron</td>
<td>4</td>
<td>(2.7%)</td>
</tr>
<tr>
<td>Prednisone</td>
<td>8</td>
<td>(2.7%)</td>
</tr>
<tr>
<td>Aspirin</td>
<td>7</td>
<td>(2.4%)</td>
</tr>
<tr>
<td>Ibuprofen</td>
<td>7</td>
<td>(2.4%)</td>
</tr>
<tr>
<td>Lamotrigine</td>
<td>5</td>
<td>(1.7%)</td>
</tr>
<tr>
<td>Loratadine</td>
<td>6</td>
<td>(2.1%)</td>
</tr>
</tbody>
</table>

Figure 1- Data flow diagram
How Can Artificial Intelligence Tools Be Used to Manage Future Pandemics? A Scoping Review of Key Use Cases

Ania Syrowatka, PhD1,2, Masha Kuznetsova, MPH3, Ava Alsabai, MA1, Adam L. Beckman, BS2,3, Paul A. Bain, PhD2, Kelly J. Thomas Craig, PhD4, Jianying Hu, PhD5, Gretchen Purcell Jackson, MD, PhD4,6, Kyu Rhee, MD, MPP4,7, David W. Bates, MD, MSc1,2

1Brigham and Women’s Hospital, Boston, MA; 2Harvard Medical School, Boston, MA; 3Harvard Business School, Boston, MA; 4IBM Watson Health, Cambridge, MA; 5IBM Research, Yorktown Heights, NY; 6Vanderbilt University Medical Center, Nashville, TN; 7CVS Health, Wellesley Hills, MA

Introduction

Coronavirus disease 2019 (COVID-19) has had a profound impact on all aspects of society, and future pandemics are sure to follow. Seven months into COVID-19, it was estimated that 59-92% of U.S. deaths could have been avoided if the pandemic were managed differently and rates reflected those in counties with moderate COVID-19 mortality.1 Policymakers, clinicians, and other stakeholders need access to data and recommendations in near-real time to respond effectively as a pandemic evolves. Artificial intelligence (AI) could be widely used to inform clinical and public health decision making. The objective of this scoping review was to identify the key use cases for involving AI for pandemic preparedness and response from the peer-reviewed, preprint, and grey literature.

Methods

We identified studies by searching five databases (PubMed, Embase, Web of Science, IEEE Xplore, ACM) and two preprint servers (medRxiv, bioRxiv), and used a structured Google search to identify grey literature. This scoping review had two parts: an in-depth review on the use of machine learning (ML) for pandemic preparedness or response, and a limited review of traditional modeling approaches. ML applications from the in-depth review were categorized into use cases related to public health and clinical practice, and narratively synthesized. Data sources and types of ML well suited for each use case were summarized. The limited review of traditional approaches served to identify additional areas where ML could be leveraged for pandemic preparedness and response.

Results

One hundred and eighty-three articles met the inclusion criteria for the in-depth review and reported on the development of ML algorithms or tools in response to COVID-19, the 2009 H1N1 pandemic influenza, severe acute respiratory syndrome (SARS), or hypothetical pandemics. The in-depth review identified six key use cases (Box 1). Data sources and types of ML that were useful varied by use case. The limited review included 1,167 articles that reported on traditional modeling approaches and identified several additional areas where ML could be applied to improve the accuracy of estimations or projections.

Discussion and Conclusion

Important ML-based solutions have been developed in response to pandemics and particularly for COVID-19. However, most were still at the research or developmental stage and had not been widely used to inform clinical or public health decisions early in the COVID-19 pandemic. Most examples of ML tools that were rapidly implemented were identified through the grey literature and were developed by health systems or industry. This review highlighted that there was still a strong reliance on traditional approaches in response to COVID-19 and identified additional areas where ML could be leveraged for pandemic preparedness and response. These findings can support stakeholders in prioritizing research and development to operationalize AI to help effectively manage future pandemics.

References

Trauma Triage in an Information Rich Environment

Douglas A. Talbert, PhD¹, Steve Talbert, RN, PhD²
¹Tennessee Tech University, Cookeville, TN; ²University of Central Florida, Orlando, FL

Introduction

Trauma research has improved triage guideline accuracy while maintaining a simplicity that respects associated time and resource constraints, but accuracy goals have yet to be achieved¹. Presuming that our increasingly sensor-rich, connected society will enable rapid capture of prehospital data that allow the use of more complex models, we relax that simplicity constraint, which can hurt understandability and reduce trust. We make this accuracy/complexity trade-off explicit to inform future work to improve understandability while maintaining accuracy improvements.

Experiments and Results

We used trauma registry data from a Level 1 Trauma Center that included physiological parameters, anatomical criteria, mechanism of injury, and age with severe injury defined by an injury severity score > 15. We then created three feature sets and multiple patient subsets. The Baseline feature set (4 features) simulated human-friendly guidelines, the Full set included all 32 available prehospital features, and the CFS set included 8 features selected using correlation-based feature selection. We compared the accuracy of models using all combinations of feature sets and datasets, but the results presented here use the Full ED dataset, which includes all patients with complete initial ED physiological values (n=56,888). All models were built using Weka’s J48 decision tree algorithm with reduced error pruning to mitigate overfitting. Accuracy metrics were measured by averaging 10 runs of 10-fold cross-validation, and for consistent comparison, we tuned all models to 95% sensitivity.

We assessed model complexity using two complexity-based XAI metrics, complexity (the number of variables in the explanation or rule) and explainability², which is defined as \[ w_1 \frac{N_i}{N_a} + w_2 + w_3 (1 - I) \], where \( N_i \) is the number of variables input into the model, \( N_a \) is the number of variables used in the explanation, \( I \) is the strength of interaction among the variables on a 0 to 1 scale. We set, \( w_1 = w_2 = w_3 = \frac{1}{3} \).

The average Full feature set specificity (57.1%) is significantly higher than that of the Baseline feature set (29.8%), and the CFS feature set’s specificity (52.3%) maintained most of the accuracy gain using a much smaller feature set. Table 1 shows that the Full set resulted in higher complexity and lower explainability than the Baseline set. The CFS set, however, resulted in the lowest complexity and an explainability between the other two sets.

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>Distinct Rules</th>
<th>Average Complexity</th>
<th>Average Explainability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>33.33</td>
<td>2.6</td>
<td>58</td>
</tr>
<tr>
<td>CFS</td>
<td>14</td>
<td>2.47</td>
<td>47</td>
</tr>
<tr>
<td>Full</td>
<td>401.33</td>
<td>7.27</td>
<td>34</td>
</tr>
</tbody>
</table>

Conclusion

We have demonstrated that a hypothetical system in a sensor-rich, connected environment can outperform other published trauma triage metrics, albeit with a corresponding increase in model complexity. Further research is needed to strengthen clinician confidence in complex models. Explanatory AI and uncertainty quantification are active areas of research that can potentially improve trust.

References

Supervised Machine Learning of Nursing Flowsheet Data to Identify a Signal of Racial Bias
Brittany N. Taylor, BS, RN1; Christopher Knaplund, MPhil2; Sarah Collins Rossetti, PhD, RN1,2; Kenrick D. Cato, PhD, RN1,3

1School of Nursing, Columbia University, New York, NY; 2Department of Biomedical Informatics, College of Physicians and Surgeons, Columbia University, New York, NY; 3Department of Emergency Medicine, New York Presbyterian Hospital, New York, NY

Introduction
Nursing documentation in electronic health records (EHR) contains information about patients’ clinical status and any changes in their condition1. The Communicating Narrative Concerns Entered by RNs (CONCERN) study linked increases in documentation frequency to increased concern for patients2. Previous studies have suggested that racial biases among nurses contribute to health care disparities3. This study represents a first analysis to determine if there is a signal for racial bias in nursing flowsheet data.

Methods
We collected nursing flowsheet annotations from the CONCERN study for the first 24 hours after admission to adult medical-surgical or intensive care units between June and July 2020. Each patient was classified as white or non-white, which included other. Using scikit-learn, we applied three supervised machine learning algorithms to train models, compare model performance, and identify important features.

Results
Of the total 4,840 visits, 1,741 were white patients, and 2,174 were non-white patients; the patients who declined to answer were excluded. CatBoost was the best performing machine learning algorithm (Table 1). Flowsheet items were highly predictive for the two different race classes; feature importance for white patients is shown on the right and non-white patients on the left. The y-axis shows flowsheet item numbers; for example, 3040109070.0 is interpreter provider service and 305610.0 is ambulation response.

Table 1. Comparison of Model Performance

<table>
<thead>
<tr>
<th>Model Type</th>
<th>Data Type</th>
<th>AUC</th>
<th>Logloss</th>
<th>PR_AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dummy Classifier</td>
<td>Train</td>
<td>0.500000</td>
<td>0.686000</td>
<td>0.720147</td>
</tr>
<tr>
<td>Dummy Classifier</td>
<td>Test</td>
<td>0.500000</td>
<td>0.691287</td>
<td>0.731163</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>Train</td>
<td>0.850918</td>
<td>0.468698</td>
<td>0.835873</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>Test</td>
<td>0.624992</td>
<td>0.890993</td>
<td>0.577486</td>
</tr>
<tr>
<td>Random Forest Classifier</td>
<td>Train</td>
<td>1.000000</td>
<td>0.194560</td>
<td>1.000000</td>
</tr>
<tr>
<td>Random Forest Classifier</td>
<td>Test</td>
<td>0.616829</td>
<td>0.669430</td>
<td>0.548661</td>
</tr>
<tr>
<td>CatBoost Classifier</td>
<td>Train</td>
<td>0.999745</td>
<td>0.278290</td>
<td>0.999656</td>
</tr>
<tr>
<td>CatBoost Classifier</td>
<td>Test</td>
<td>0.650018</td>
<td>0.656527</td>
<td>0.615000</td>
</tr>
</tbody>
</table>

Conclusion
We found a signal in the data, indicating the presence of a racial bias in nursing documentation. Analysis of these data is ongoing, and results based on comorbidity and discharge diagnosis have not been completed and will be presented in the poster at the annual meeting. Additional research is needed to analyze nursing documentation patterns by setting, since previous research by our study team has detected significant differences in intensive care versus medical/surgical inpatient wards1.

Funding: This work was supported by the NIH/NINR under award numbers T32NR007969-04S1 and R01NR016941 and the NLM under award numbers T15LM007079 and RO1LM06910.

References
Telehealth Adoption and Healthcare Utilization During the COVID-19 Pandemic: Toward Understanding What Follows for Vulnerable Groups

Casey Overby Taylor, PhD; Ilia Rattsev, BS; Jeremy Epstein, MD; Jiawei Bai, PhD; Natalie Flaks Manov, PhD

1The Johns Hopkins University, Baltimore, Maryland, United States

Abstract: The COVID-19 pandemic provides a natural experiment to compare healthcare utilization among different demographic and community groups with the broad introduction of telehealth. Our findings suggest that individuals with high healthcare utilization pre-pandemic maintained a similar rate of visits during the pandemic, and that a large proportion of most groups compensated for declines in face-to-face visits with telehealth.

Objective: The objective of this study was to investigate the dynamics of telehealth adoption among Johns Hopkins patients with high healthcare utilization prior to the pandemic, by socio-economic and demographic characteristics.

Materials and Methods: This was a retrospective study leveraging electronic health record data of patients that had a Johns Hopkins Medicine (JHM) primary care provider visit at least twice within twelve months prior to the onset of the pandemic (Mar-Dec 2019). Encounter types included completed outpatient visits (face-to-face and telehealth), four demographic characteristics (age, gender, race, and ethnicity); and four socio-economic community factors (poverty, education, unemployment, and vehicle possession). The demographic factors were obtained for each patient from Epic, while the community characteristics were assigned to patients according to neighborhood-level US Census data. We analyzed the trends of outpatient visits by month for each encounter type. We also assessed rates of visits pre-pandemic and pandemic (i.e., total visits in a period over the total population visits during the same period.) To adjust for seasonal fluctuations in the dynamics of interaction with healthcare, the inclusion window timeframe for our analysis covered respective months pre-pandemic (Jun 1 to Dec 31, 2019) and pandemic (Jun 1 - Dec 31, 2020).

Results: The cohort included a total of 93,980 patients. 29,784 patients (32%) were 65 years old or older, the majority were White or Caucasian (55%, 51,745), with Black or African American being the second most represented race (31%, 29,229). There were 637,307 completed outpatient visits (both face-to-face and telehealth) and 336,529 portal messages in the pre-pandemic period. In the pandemic period, these numbers were 479,255 for outpatient visits. Findings from reviewing visit trends showed that telehealth visits were rarely used prior to the pandemic, and the rapid increase in telehealth use happened at the onset of the pandemic (Mar 2020, Figure 1). Face-to-face visits did not stop, but occurred to a lesser extent during the pandemic, while the total amount of visits has dropped (Mar 2020 to Dec 2020). The overall rate of visits was 4.9 pre-pandemic and 4.7 during the pandemic. For two groups, this drop approached 1 fewer visits on average per person: age 0-24 (3.5 to 2.7) and >= 40% in extreme poverty (7.7 to 7.0).

For groups maintaining a more similar rate of visits pre- to pandemic, over 50% of people had at least one telehealth visit (seven groups with 60% or more: Females, age 25-44, age 45-64, Black or African American, < 75% with high school diploma, >= 40% in extreme poverty, > 20% unemployed, >20% with no vehicle).

Discussion: Trends in healthcare utilization pre-pandemic and pandemic were similar to the findings of others. Most groups maintained a similar rate of visits during the pandemic as pre-pandemic. There was also a high proportion of people with at least one telehealth visit during the pandemic, including for some vulnerable groups.

Figure 1. Trends of outpatient patient-provider communications by month for two encounter types (face-to-face visits and telehealth visits [phone and video]).

Towards building a Recommender System: Analyzing Resource Sharing on an Online Ovarian Cancer Community

Khushboo Thaker\textsuperscript{a}, Susan Birkhoff\textsuperscript{b}, Vivian Hui\textsuperscript{c}, Young Ji Lee\textsuperscript{c}, Peter Brusilovsky\textsuperscript{a}, Daqing He\textsuperscript{a}

\textsuperscript{a}. Intelligent Systems Program, School of Computing and Information, University of Pittsburgh, USA
\textsuperscript{b}. Department of Acute & Tertiary Care, School of Nursing, University of Pittsburgh, USA
\textsuperscript{c}. Department of Health and Community Systems, School of Nursing, University of Pittsburgh, USA

Introduction: Ovarian cancer (OvCa) can be a deadly gynecological cancer affecting about 22k women per year in the United States with a high recurrence rate [1]. Women with OvCa and their caregivers often seek support from online health communities (OHCs) [2]. These OHCs allow for the exchange of information and resources with other individuals who have had similar experiences. Prior research studied OvCa patients’ informational needs using OHCs [2], yet there is a paucity of research investigating the types of resources shared among users, such as articles or websites. The overarching goal of our project (R01LM013038) will be to develop a recommender system that personalizes online health-related materials to patients with OvCa and their caregivers. As a part of this project, we studied the types of resources shared by users on an online OvCa OHC to recognize health materials sought after by patients with OvCa and their caregivers.

Data and Methodology: Data extracted from the National Ovarian Cancer Coalition online forum contained 909 threads. Each thread included anonymized user details, thread title, initial post, and reply comments. We selected 105 initial threads based on users asking for explicit information need from other users. Within these initial threads, users shared Uniform Resource Locators (URLs) in their reply comments. These URLs were extracted for analysis utilizing regular expression [3]. Manual content analysis was performed for each URL and these URLs were classified based on their content (health blog, health news, research articles, patient guidelines and factsheets).

Results: Of the 176 URLs shared among users, 112 were direct links to reading materials. The 64 other URLs belonged to events, organization websites, doctor listings, videos, and shopping sites. Users primarily exchanged commercial, government, health organization and academic sites (.com, .gov, .edu and .org) (Figure 1). The most shared domains were news.cancerconnect.com, cancerconnect.com, nccn.org and harvard.edu. Classification of 112 URLs, as shown in Figure 2 revealed, OvCa OHC users predominately exchanged health news and health blogs, which summarized research and clinical trial findings rather than sharing original research articles. Infrequently, research articles were shared eleven times within reply to threads.

Conclusion: Analyzing these exchanges of information among users on an OvCa online forum builds the preliminary evidence demonstrating OvCa OHC users primarily rely on articles which summarizes the medical findings in layman terms. This insight informs our HELPeR recommender system to prioritize these shared resources over research articles.

References

Assessment of Fast Healthcare Interoperability Resources Response Times in a Production Electronic Health Record Embedded Visualization

Jeritt G. Thayer1, Megan O. Lewis MSN CRNP1, Jonathan Spergel MD PHD1,2, Robert W. Grundmeier MD1,2
1Children’s Hospital of Philadelphia, Philadelphia, PA; 2Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA

Introduction
Clinical decision support (CDS) has demonstrated the ability to improve patient safety and the quality of patient care. These systems are often developed using internal electronic health record (EHR) rules engines that collect patient information directly from the chart and provide alerts (e.g. drug allergies) or reminders (e.g. preventive healthcare interventions such as vaccine administration) to the care provider. With recent advancements in health information exchange, such as the introduction of Fast Healthcare Interoperability Resources (FHIR),1 health organizations are now able to more easily take advantage of external CDS services, which often help to provide more robust visualization of patient information or computation of complex decision logic. Unfortunately, the use of FHIR for CDS applications has traditionally been limited to small timeframes (e.g. within the same encounter) or for retrieving specific values from the patient’s chart (e.g. current medications). Limited information exists about the effectiveness of FHIR application programming interfaces (API) for aggregating data across a larger time frame in the context of clinical care. At the Children’s Hospital of Philadelphia (CHOP), we sought to assess the capabilities of FHIR APIs in the context of an EHR-integrated production information visualization.

Methods
We utilized application performance monitoring logs captured between October 26, 2020 and December 10, 2020 from a single EHR-integrated production application used by CHOP’s Division of Allergy and Immunology outpatient clinics. The application is an information visualization of a patient’s asthma history, which presents two years of relevant asthma data. The application utilizes four Release 4 (R4) FHIR application programming interfaces (API) – Patient, Encounter, Procedure, and Condition. Details of each API request, which includes how long the EHR web server took to respond, was logged by the application to our hospital logging infrastructure. We also logged the total amount of time it took to load the application, which, in addition to requesting data, includes activities such as executing clinical logic and building the visualization. We evaluated each FHIR resource and the total application load time based on four response time percentiles (50th, 75th, 90th, and 95th), which are common thresholds for monitoring application response times.

Results
During our analysis period, the application was used by 124 users for 2,396 patients in 3,790 encounters. Average application response time across all patients was 1.052 seconds. Average FHIR API response time was: Patient – 0.187 seconds; Encounter – 0.635 seconds; Procedure – 0.204 seconds; and Condition – 0.169 seconds. Our analysis showed that, across all percentiles, the FHIR Encounter resource was the slowest service to respond (Table 1). This was especially pronounced for patients deemed “high utilizers” by the application, where 5% of requests took more than 5 seconds to respond.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>50% Low</th>
<th>75% Medium</th>
<th>90% High</th>
<th>95% Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Application</td>
<td>0.765</td>
<td>1.172</td>
<td>1.801</td>
<td>2.408</td>
</tr>
<tr>
<td>R4 FHIR Resource</td>
<td>1.198</td>
<td>1.699</td>
<td>2.393</td>
<td>3.004</td>
</tr>
<tr>
<td>Patient</td>
<td>0.116</td>
<td>0.219</td>
<td>0.377</td>
<td>0.534</td>
</tr>
<tr>
<td>Encounter</td>
<td>0.386</td>
<td>0.678</td>
<td>1.613</td>
<td>2.330</td>
</tr>
<tr>
<td>Procedure</td>
<td>0.137</td>
<td>0.217</td>
<td>0.399</td>
<td>0.581</td>
</tr>
<tr>
<td>Condition</td>
<td>0.019</td>
<td>0.188</td>
<td>0.303</td>
<td>0.440</td>
</tr>
</tbody>
</table>

Table 1 Response time percentiles measured in seconds partitioned by healthcare utilization for asthma
* Utilization level was determined by the application.

Discussion
We explored the real-world use of FHIR in the context of a production clinical application. Our work demonstrates that requests for data using the FHIR standard, as implemented by one commercial EHR, may be unacceptably slow to be useful in typical clinical workflows. Previous research has suggested that users’ thought processes can be interrupted for response times greater than 1 second.[2] Our study was limited to a single institution and the analysis does not account for potential variation in the underlying information architecture (e.g. server load, network latency, EHR database). Future work should focus on assessing the impact of long application load times as well as developing methods to reduce such impacts.

References
Vancomycin Dosing in Critically Ill Patients: A Machine Learning Approach

Mohammad Samie Tootooni, Ph.D.\textsuperscript{1}, Erin F. Barreto, Phram.D., M.Sc.\textsuperscript{2}, Kalyan S. Pasupathy, Ph.D.\textsuperscript{2}, Kianoush B. Kashani, M.D., M.Sc.\textsuperscript{2}

\textsuperscript{1}Loyola University Chicago, Maywood, IL, USA; \textsuperscript{2}Mayo Clinic, Rochester, MN, USA

As a renally-eliminated and nephrotoxic medication, both sub- and supra-therapeutic vancomycin trough concentrations have consequences. A high vancomycin level can lead to drug-associated nephrotoxicity, whereas low levels could result in therapeutic failure and antimicrobial resistance. Selecting the right drug dose and frequency remains challenging, particularly for complex patients such as those in the intensive care unit (ICU) \cite{1}. Despite several efforts to optimize dosing methods, it is recently demonstrated that the majority of critically ill patients do not reach the targeted trough concentration until at least three days into their course \cite{2}. We aimed to identify the key predictive factors for the vancomycin steady-state trough level and their relative contribution, estimate the risk of a steady-state trough outside the goal range, and recommend the optimal individualized dosing recommendations.

A theoretical timeline for vancomycin concentration in a patient's serum is shown by Figure 1. At each decision point (DP), we aimed to have two classification models to accurately predict the class within which the steady-state trough concentration would fall (i.e., sub-therapeutic, therapeutic, and supra-therapeutic). Five-fold cross-validation was used for training and validation. We trained and tested six models selected partially based on maximum dissimilarity between model structures. The prediction models with highest performance are used to develop our recommendation engine. To find the optimum recommendation, a discrete optimization problem was formulated where the decision variables are: Loading Dose ($LD$), Maintenance Dose ($MD$), and Interval ($T$).

Our cohort consists of adult patients admitted to the ICUs at Mayo Clinic, Rochester, MN. All ICU patients from 2007 to 2017, who had a vancomycin concentration measured were included. The final cohort consisted of 5,337 patients, 80% used for training and cross-validation purposes, and 20% were hold-out (test set) for performance measurement. The xgbTree ensemble models were finally chosen to be tested via the left-out set in predicting sub-therapeutic (ROC: 0.85, Specificity: 0.53, and Sensitivity: 0.94) and supra-therapeutic (ROC: 0.83, Specificity: 0.47, and Sensitivity: 0.94) categories, respectively. Also, the Hosmer and Lemeshow calibration tests were performed, which suggested that the output probabilities from the xgbTree models embodied the actual risk of the defined therapeutic categories without additional calibration steps. The serum creatinine related measurements, prior dosages and frequencies, comorbidities, BMI, age, gender, and Mean Arterial Pressure were among the most contributing variables.

The optimal recommendations were calculated and compared with historical dosing decisions to measure the potential improvement in practice in a retrospective setting. McNemar’s tests were performed to measure the potential effectiveness from using the recommendations and shown in Table 1. The $P$ values ($<0.0001$) showed the outcome improvement by following the recommendations were statistically significant. Details showed the recommendations are more accurate for detecting/preventing supra-therapeutics compared to sub-therapeutic. The next steps include additional efforts for improving the recommendations and retrospective and prospective validations.

Extracting Daily Dosage from Medication Instructions in EHRs: An Automated Approach and Lessons Learned

Ching-Huei Tsou, PhD, Diwakar Mahajan, MS, Jennifer J. Liang, MD
IBM TJ Watson Research Center, Yorktown Heights, NY

Introduction
Medication timelines have been shown to be effective in helping physicians visualize complex patient medication information. A key feature in such designs is a longitudinal representation of a medication’s daily dosage and its changes over time. However, daily dosage as a discrete value is generally not provided and needs to be derived from free text instructions (Sig). Existing works in daily dosage extraction are narrow in scope, targeting dosage extraction for a single drug from clinical notes. Here, we present an automated approach to calculate daily dosage for all medications, combining deep learning-based named entity extractor with lexicon dictionaries and regular expressions.

Methods
We define the task as a 2-step process. First, we employed ClinicalBERT to train a model to extract two entities of interest from the Sig: DosagePerAdministration and AdministrationFrequency. To overcome lack of data, we adapted annotations from the publicly available n2c2 2018 Adverse Drug Events and Medication Extraction in EHRs dataset to form our entities of interest. Second, we normalized extracted spans to corresponding numerical values using a pre-built lexicon, and then applied a set of rules on the normalized values to calculate the daily dosage. Figure 1 presents an overview of our system. For evaluation, we sampled 1000 medication orders from 427 patients within a large multidisciplinary medical center for ground truth generation. Annotators were provided with the Sig text (“1-2 tabs qd”) and medication strength (“7.5mg”) to calculate the minimum (“7.5mg”) and maximum (“15mg”) daily dosage.

Results and Discussion
Out of 1000 Sigs, human experts were able to provide a daily dosage in 83% of the data. The remaining 17% consists of cases with missing or contradicting instructions, variable dosing over different days, cases where daily dosage is not meaningful (e.g. non-routine dose, single administration), and cases where dosage is not easily quantifiable (e.g. topical creams, eye drops). When evaluated on this dataset, our system achieved 0.98 precision, 0.95 recall, 0.96 F-score and 96.2% accuracy. Another way to understand these results is by comparison to human experts. In our evaluation dataset, human experts provided a daily dosage in 83% of the data and is 100% correct when a daily dosage value is returned, whereas our system returned a daily dosage in 81.6% of the data and is 98.2% correct when returned. Error analysis of our system revealed 4 main categories of errors: multiple dosage expressions, new expressions not understood by system, extra information in Sig not understood by system, and typos/misspellings.

Conclusion
To the best of our knowledge, this work is the first to generalize the task of daily dosage extraction on all medications. We hope this research will help realize many medication timeline designs and ultimately reduce the cognitive burden on healthcare providers. In future, we plan to improve our system by developing a generalizable end-to-end seq2seq model to calculate daily dosage and thereby removing the reliance on the manually curated lexicon.

References
Inconsistencies in handling missing data across stages of prediction modelling: a review of methods used

Antonia Tsvetanova, BSc MRes, PhD¹, Matthew Sperrin, PhD¹, Niels Peek, PhD², Iain Buchan³, MD, FFPH, FACMI, Stephanie Hyland⁴, PhD, Glen Martin, PhD¹
¹Centre for Health Informatics, Faculty of Biology, Medicine and Health, University of Manchester, UK; ²NIHR Manchester Biomedical Research Centre, Faculty of Biology, Medicine and Health, University of Manchester, Manchester Academic Health Science Centre, Manchester, UK; ³Microsoft Research Cambridge, UK; ⁴Institute of Population Health, The University of Liverpool

Introduction

Missing data need to be addressed at each stage of developing, validating and implementing a clinical prediction model (CPM). No clear guidance exists on handling missing data across this pipeline and it is unknown which methods are used in practice. We aimed to review the approaches to handling missing data that underly the CPMs currently recommended for use in UK healthcare.

Methods

We identified eligible CPMs through discussions with National Institute for Health and Care Excellence (NICE), a call on Twitter, and contacting other research groups. We identified the paper corresponding to each model’s development and identified the ten most cited external validation papers. We extracted information on methods used for handling missing data, as well as reported strengths and limitations, and any stated assumptions. We preferred to get information on the implementation stage for each CPM from documentation provided by the CPMs’ developers. If this was not available, we utilised online calculators to infer how missing data were handled during the implementation stage.

Results

23 CPMs were included. Six missing data strategies were identified: complete case analysis (CCA), multiple imputation, imputation of mean values, k-nearest neighbours imputation, using an additional category for missingness, considering missing values as risk-factor-absent, as shown on Figure 1. 52% of the development articles and 48% of the validation articles did not report how missing data were handled.

Conclusion

Missing data handling strategies were generally inconsistent. Our future work will involve a simulation study to explore whether these inconsistencies have an effect on the predictive performance of CPMs.
Comparison of Three Phrase Chunking Approaches on Medical Concept Coverage in Clinical Text

Grace K. Turner, BSE. and Meliha Yetisgen, PhD.
Biomedical Informatics and Medical Education, University of Washington, Seattle, WA

Abstract

Medical concept extraction pipelines commonly first chunk text into phrases (sets of words and punctuation such as “M.I.” or “myocardial infarction”) and then map those phrases to controlled vocabularies. In this work, we explored the effect of different phrase chunking approaches on medical concept coverage by using a subset of the i2b2 2010 medical concept dataset\(^1\) as the gold standard. By selecting the best chunking mechanism, we can potentially improve the entire pipeline performance and better facilitate access to the information stored in clinical notes.

Methods and Discussion

In our experiments, we used the official test subset of the i2b2 2010 dataset, with 45004 unique annotations over 477 reports. Including stop words, the number of tokens or words per annotation ranges between 1 and 26 with an average of 2.16. We evaluated the following three types of chunking approaches. (1) N-grams: We chunked text into basic n-grams (1,2,3-gram). (2) Stanza: We used Stanford’s Stanza CoreNLP package\(^2\) to generate all possible noun phrases (NP) and verb phrases (VP) based on part of speech tags. (3) Wildgram: We developed a heuristic approach called Wildgram that chunks text based on at least two instances of white spaces, punctuation marks, and stop words. A sentence such as “the patient had a myocardial infarction” would therefore be split into two tokens: “patient” and “myocardial infarction”. We also evaluated two combinatorial methods. The first combined the phrases created from Wildgram and 1-gram. The second combined the phrases from 1-gram, 2-gram, and 3-gram. We measured the quality of the chunking methods through precision, recall, and the F-1 score of perfect span overlap. Additionally, we measured the speed and the total number of chunks that partially match the true annotations.

Table 1: Concept coverage performance of different chunking methods on the 2010 i2b2 dataset

<table>
<thead>
<tr>
<th>Method</th>
<th># of Overlapping Chunks</th>
<th>Precision</th>
<th>Recall</th>
<th>F-1</th>
<th>Avg Speed (Seconds per 100 Chars)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-gram</td>
<td>81689</td>
<td>0.242</td>
<td>0.44</td>
<td>0.313</td>
<td>0.00135</td>
</tr>
<tr>
<td>2-gram</td>
<td>126685</td>
<td>0.079</td>
<td>0.223</td>
<td>0.117</td>
<td>0.00134</td>
</tr>
<tr>
<td>3-gram</td>
<td>171681</td>
<td>0.025</td>
<td>0.095</td>
<td>0.039</td>
<td>0.00134</td>
</tr>
<tr>
<td>Stanza</td>
<td>126563</td>
<td>0.206</td>
<td>0.584</td>
<td>0.305</td>
<td>0.668</td>
</tr>
<tr>
<td>Wildgram</td>
<td>61965</td>
<td>0.45</td>
<td>0.62</td>
<td>0.522</td>
<td>0.0016</td>
</tr>
<tr>
<td>Wildgram and 1-gram</td>
<td>127754</td>
<td>0.248</td>
<td>0.704</td>
<td>0.367</td>
<td>0.003</td>
</tr>
<tr>
<td>1,2,3-gram</td>
<td>380055</td>
<td>0.09</td>
<td><strong>0.758</strong></td>
<td>0.161</td>
<td>0.0042</td>
</tr>
</tbody>
</table>

Table 1 presents the concept coverage performance on the 2010 i2b2 test subset. Wildgram had the highest precision and F-1 score overall. Combinatorial methods resulted in higher recall with a consequent reduction in precision due to the significant increase in the total number of generated spans. Wildgram and n-gram methods ran at around the same speed, while Stanza is significantly slower. N-gram methods typically failed on phrases of a different size. For example, the 1-gram method failed on two word phrases. Wildgram typically failed most with sentences that had fewer stop words or other natural breakpoints. Wildgram can be downloaded as a python package.

References

Advance Care Planning Documents Should be ‘FAIR’: Towards Findable, Accessible, Interoperable, and Reusable Advance Care Planning Documentation

Elizabeth Umberfield, PhD, RN\textsuperscript{1,2}, Susan E. Hickman, PhD\textsuperscript{1,3,4}, Titus K. Schleyer, DMD, PhD, FACMI, FAMIA\textsuperscript{1,4}

\textsuperscript{1}Regenstrief Institute, Indianapolis, IN; \textsuperscript{2}Fairbanks School of Public Health, Indiana University, Indianapolis, IN; \textsuperscript{3}School of Nursing, Indiana University, Indianapolis, IN; \textsuperscript{4}School of Medicine, Indiana University, Indianapolis, IN

Introduction

It is ethically imperative that providers honor patients’ end of life treatment preferences. However, advance care planning (ACP) documentation often exists solely on paper or is buried in the electronic health record (EHR).\textsuperscript{1} A lack of standardized and interoperable strategies for storing ACP documents make it challenging to find ACP documents to guide treatment decisions; this increases the likelihood preferences will not be honored. This review examines challenges for exchanging ACP documentation and proposes generalizable solutions derived from the framework of the FAIR Guiding Principles (i.e., Findable, Accessible, Interoperable, and Reusable), which was originally proposed for management of scientific data rather than clinical documents.\textsuperscript{2}

Methods

This derivation of the FAIR Principles follows review of the literature and semi-structured qualitative interviews with clinicians to understand current approaches for and challenges when accessing advance care plans through the electronic health record. We comprehensively examine challenges for ACP document exchange across a range of settings and perspectives, including: end users; health information management departments; and a regional health information organization (RHIO). We derive and adapt each of the FAIR Principles,\textsuperscript{3} translating them for ACP document exchange rather than research data management.

Results

ACP document interoperability remains underdeveloped. Despite multiple bespoke and proprietary tools to facilitate ACP documentation in electronic workflows, few enable exchange across information systems. Applied to ACP, the FAIR principles could be initially adapted as follows:

- **F** ACP documents must be indexed in a searchable resource.
- **A** ACP document access must be open to providers, contingent upon robust authorization and authentication.
- **I** ACP common data elements and their value sets must be based on broadly accepted knowledge representation languages and standardized vocabularies.
- **R** ACP documents must be linked with detailed provenance records, demonstrating document revision and approval over time.

Conclusion

The FAIR Principles provide a solid framework for HIT infrastructure which may support ACP document exchange and discovery. The vision for seamless health information exchange has the greatest likelihood of being realized when it is supported by rich data and metadata annotation. This will ensure information is discoverable and actionable by its intended users and contribute to the provision of dignifying care for patients towards end of life.

References

DEVELOPING CLINICAL PREDICTION MODELS USING PRIMARY CARE ELECTRONIC HEALTH RECORD DATA – THE IMPACT OF METHODOLOGICAL CHOICES ON MODEL PERFORMANCE

Hendrikus J. A. van Os, MD*1,2,3; Jos P. Kanning, MSc*4; Marieke J. H. Wermer, MD, PhD2; Niels H. Chavannes, MD, PhD1,2; Mattijs E. Numans, MD, PhD3; Ynte M. Ruigrok, MD, PhD4; Erik van Zwet, PhD5; Hein Putter, PhD2; Ewout W. Steyerberg, PhD2; Rolf H. H. Groenwold, MD, PhD5,6 *Shared first authors

1Department of Neurology, 2National eHealth Living Lab, 3Department of Public Health & Primary Care, Leiden University Medical Hospital, 4Department of Neurology, University Medical Center Utrecht, 5Department of Biomedical Data Sciences, 6Department of Clinical Epidemiology, Leiden University Medical Hospital.

Problem description: Earlier research already identified several methodological challenges when developing clinical risk prediction models using electronic health record (EHR) data.1 This study quantifies the impact of these challenging methodological choices on model performance.

Study Design and Setting: Cox proportional hazards models were developed predicting cardiovascular events using Dutch primary care EHR data. The reference model was based on a one-year run-in period, cardiovascular events were defined based on both EHR diagnosis and medication codes, and missing values were multiply imputed. We compared methodological choices regarding i) length of the run-in period (more years); ii) outcome definition (EHR diagnosis codes or medication codes only); and iii) methods addressing missing values (mean imputation and complete case analysis) by making variations on the derivation set and testing their impact in a validation set.

Results: We included 89,491 patients in whom 6,736 cardiovascular events occurred during a median follow-up of 8 years. Outcome definition based only on diagnosis codes led to systematic underestimation of risk (intercept:0.835;95%CI:0.827–0.843), while complete case analysis led to overestimation (intercept:0.519;95%CI:0.509–0.529). Differences in length of run-in period showed no relevant impact on calibration and discrimination.

Conclusion: Methodological choices regarding outcome definition or methods addressing missing values can have a substantial impact on the calibration of predictions, hampering reliable clinical decision support. Transparent reporting of modelling choices is hence essential.

<table>
<thead>
<tr>
<th>Methodological challenge</th>
<th>Derivation set description</th>
<th>C-statistic (95% CI)</th>
<th>Calibration curve intercept (95% CI)</th>
<th>Calibration curve slope (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference model*</td>
<td>NA</td>
<td>0.670 (0.669 - 0.671)</td>
<td>-0.004 (-0.012 - 0.003)</td>
<td>0.996 (0.989 - 1.003)</td>
</tr>
<tr>
<td>Run-in variations</td>
<td>2 years run-in</td>
<td>0.668 (0.667 - 0.67)</td>
<td>0.074 (0.068 - 0.081)</td>
<td>1.078 (1.071 - 1.084)</td>
</tr>
<tr>
<td></td>
<td>3 years run-in</td>
<td>0.667 (0.666 - 0.669)</td>
<td>0.159 (0.152 - 0.166)</td>
<td>1.173 (1.164 - 1.181)</td>
</tr>
<tr>
<td>Variations in definition</td>
<td>ATC (excl. ASA) or ICPC</td>
<td>0.666 (0.664 - 0.667)</td>
<td>0.401 (0.394 - 0.407)</td>
<td>1.494 (1.484 - 1.503)</td>
</tr>
<tr>
<td></td>
<td>ATC only</td>
<td>0.671 (0.67 - 0.673)</td>
<td>0.005 (-0.003 - 0.013)</td>
<td>1.005 (0.997 - 1.013)</td>
</tr>
<tr>
<td></td>
<td>ATC (excl. ASA) only</td>
<td>0.667 (0.666 - 0.668)</td>
<td>0.521 (0.513 - 0.529)</td>
<td>1.685 (1.671 - 1.699)</td>
</tr>
<tr>
<td></td>
<td>ICPC only</td>
<td>0.665 (0.663 - 0.666)</td>
<td>0.835 (0.827 - 0.843)</td>
<td>2.305 (2.288 - 2.323)</td>
</tr>
<tr>
<td>Missing data method</td>
<td>Complete Case Analysis</td>
<td>0.666 (0.665 - 0.668)</td>
<td>-0.519 (-0.529 - -0.509)</td>
<td>0.595 (0.59 - 0.601)</td>
</tr>
<tr>
<td></td>
<td>Mean imputation</td>
<td>0.666 (0.664 - 0.667)</td>
<td>-0.001 (-0.009 - 0.008)</td>
<td>1.000 (0.991 - 1.009)</td>
</tr>
</tbody>
</table>

Table 2. Performance of the models based on derivation set variations compared with the reference model in Dutch primary care EHR (n=89,491). ASA = acetylsalicylic acid; ICPC = International Classification of Primary Care diagnosis codes; ATC = Anatomical Therapeutic Chemical medication codes

References
Quantifying Uncertainty in Patient Count Metrics Derived from Imperfect EHR-Based Phenotypes
Jack Vanschaik1, Sunandan Chakraborty1
1Indiana University Purdue University of Indianapolis, School of Informatics and Computing

Cohort identification via electronic health record (EHR) data has played an increasingly important role in retrospective clinical studies1 and public health reporting. Patient cohorts are identified via EHR-based phenotypes, a broad class of algorithms that attempt to classify patient inclusion using their medical records. While EHR-based phenotypes span various levels of sophistication, from rule-based methods to deep learning algorithms, a universal issue is their imperfect accuracy2. Phenotypes are subject to classification error, which is commonly measured by sensitivity and specificity. Due to the extensive time and expert knowledge required for phenotype validation, such measures are usually unknown3. Furthermore, EHR data quality issues may lead to irreducible error in cohort validation.

Problem
Retrospective analysis of EHR data and the need for public health reporting may necessitate the use of EHR-based phenotypes, even if those phenotypes have not undergone complete validation. Desired cohort metrics, such as population sizes and relative risk ratios, are associated with uncertainty due to phenotype error. These metrics are often reported as single values or reported with confidence intervals that do not account for the unavoidable issue of phenotype accuracy.

Solution
We introduce a Bayesian model for quantifying uncertainty in patient count metrics derived from one or two imperfect EHR-based phenotypes. We account for phenotype uncertainty by placing Beta priors over phenotype sensitivity and specificity. Hyperparameters can be determined from the best-known information about phenotype performance. For example, a phenotype estimated to have sensitivity 90% +/- 5% could be modeled as a Beta distribution with α = 36 and β = 4 which has mean 0.9 and standard deviation ~0.05. Combining prior information with collected EHR data allows for sampling of a posterior distribution. This distribution can be used to create posterior estimates and confidence intervals for EHR derived cohort metrics. The model was tested using a cohort of 124,150 COVID-19 positive patients from the synthetic EHR dataset SyntheticMass4. We used the model to estimate uncertainty in the relative risk of death for Black and Hispanic patients with COVID-19. Hyperparameter selection and the resulting posterior intervals are detailed in Figure 1.

Conclusion
The model produces confidence intervals for population counts that are scaled and shifted in accordance information about phenotype uncertainty. Accuracy priors modeling perfect phenotypes (e.g. α > 1000, β = 1 for sensitivity and specificity) yield nearly identical results to frequentists estimates. Tests with synthetic data showed substantial adjustment in confidence intervals when the model accounted for phenotype imperfections. Such differences could be significant for public health reporting and research. This model can save expert reviewer hours by enabling the use of EHR-based phenotypes when they have not undergone complete gold-standard validation, or the EHR in question has data quality issues. Code for implementation of the model in the R programming language is open sourced and freely available via GitHub.

References
Integrating Hospital Compare Data into the Health Services Research Data Warehouse

Madison Vinson, MSHI, Bunyamin Ozaydin, MSEE, PhD, Ferhat Zengul, PhD
University of Alabama at Birmingham, Birmingham, AL, USA

Introduction
As performance and quality improvement healthcare initiatives are increasing, data from Hospital Compare are being extensively utilized by health services researchers. These publicly available data are gathered through the voluntary Hospital Inpatient Quality Reporting Program (IQR) and Hospital Outpatient Quality Reporting (OQR) Program and reported by The Centers for Medicare and Medicaid (CMS) (1). Hospital Compare ratings have been shown to impact financial value, influence perception, and highlight areas of improvement within hospitals (2). The current extract-transform-load (ETL) process for retrieving Hospital Compare data from the website is tedious, repetitive, and leaves room for human error.

By creating an automated ETL process using SQL Server Integration Services to incorporate and update Hospital Compare data to a health services research data warehouse that includes data from other sources as well, researchers can eliminate separate manual ETL processes for each research project. This poster will present a process to allow researchers to access previous and current Hospital Compare data without having to spend their time on data management tasks each time an analysis needs to be performed.

Methods
The process for this project is a five-step ETL process as seen in figure 1. Phase one involves separating the raw data from the Hospital Compare websites into folders based on reported date and relevant measures. Every time there is a change in the structure of a data table (column was renamed, removed, added, or added files, etc.) we consider a new version of that table. The second phase is an SSIS package consisting of one large sequence container broken into hospital, national, and state sequence containers that are then further broken into foreach loops that will hold one version of a file. The foreach loops are used to run multiple Hospital Compare files in one data flow task. Stage three is an SSIS package that works dynamically to sort through Access files, pull relevant tables, and insert them into raw tables in SQL Server. This step is done to all Hospital Compare data prior to 2013. The fourth phase of this project involved creating SQL scripts to move the data from the raw tables into relational database tables. The scripts created in this phase remove data replication and convert data into a usable format for data visualization purposes (ex. dates). The final phase of this project handles updates to Hospital Compare data. This package does not contain foreach loops like the previous archive load package because it was created to only process one release and requires little to no technical knowledge to maintain.

Results and Conclusion
The measure of this study’s effectiveness was how well the framework integrated within The University of Alabama at Birmingham’s health services research data warehouse. Numerous visualizations were made in Power BI and Tableau to sort specific measures by date, performance scores, states, cities, or hospital using the tables in the health research database. The result of this project will lessen the burden on researchers by utilizing a data warehouse and SQL server business intelligence tools to streamline the data management stage of research. The final product provides researchers with the tools to automatically process all archived files and new releases with a set of SSIS packages and SQL scripts.

References

1860
INTRODUCTION. Much like other facets of healthcare, data interoperability and data exchange are a major concern in clinical research informatics (CRI). REDCap is one of the leading CRI data collection tools and is integral to the success of many observational studies. The Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) serves as a solution for organizations across the globe to conduct systematic analysis of disparate databases. Health Level Seven (HL7) Fast Healthcare Interoperability Resources (FHIR) is a well-trusted standard for facilitating data-exchange at the electronic health record (EHR) level. There have been recent efforts to combine each of these resources: REDCap and OMOP; REDCap and FHIR; and OMOPonFHIR and more recently, an official partnership between HL7 and the Observational Health Data Sciences and Informatics (OHDSI) network who maintains OMOP. To our knowledge, no prior work offers a direct pipeline for exchanging OMOP and FHIR compliant REDCap data. As such, we are developing REDHot OMOP.

METHODOLOGY. REDHot OMOP is a SMART on FHIR application and conduit in a system that will leverage FHIR and OMOP to exchange REDCap data standardized in a format tailored toward improving translational and observational research endeavors.

The pipeline works as follows: (1) users enter information into a REDCap form with any free-text input being (1a) captured in a specialized field to a local FHIR Terminology Service that (1b) returns a list of potential codes for the user based on a pre-defined list of values (FHIR ValueSet Resource) from a local code system selected by the form designer in advance. Once completed, the form report is (2) exported into Redmatch, which (3) converts the REDCap data into FHIR Resources and sends them to REDHot OMOP. The FHIR Resources are stored in REDHot OMOP and the local codes within the Resources are (4a) sent to OMOPonFHIR, which maps the codes in the FHIR Resources to OMOP and returns all viable mappings to REDHot OMOP. REDHot OMOP then (5a) provides the user with a representation of the transformed REDCap data for uses such as visualization or covariate analysis. REDHot OMOP can (5b) connect with an EHR or external application to assist in clinical decision support or be integrated into a mobile app. Information can also be (6) extracted from the EHR and associated with a participant in future research.

DISCUSSION & CONCLUSION. REDHot OMOP will support interoperable data exchange between observational research and clinical practice and has the potential to foster robust learning health systems. Some key limitations to implementing REDHot OMOP include relying on open-source software and requiring that institutions have a local OMOP instance and FHIR terminology server for concept mapping. This work could have lasting implications on CRI, observational and translational research.

Acknowledgements. The authors thank George Costakis, Weewei Xu, Li Li, Zoe Jiang, and Taha Abdul-Basser from NYPH, Myung Choi and Elizabeth Shivers from Georgia Tech, and Alejandro Metke-Jimenez from CSIRO for their assistance.

References
2. HL7 FHIR, https://www.hl7.org/fhir/
5. OMOPonFHIR, http://omoponfhir.org/
6. OHDSI and HL7 partnership, https://www.ohdsi.org/ohdsi-hl7-collaboration
7. REDHot OMOP, https://github.com/salvolpe/REDHotOMOP
**DS-DETERMINED: Using PCORnet and Referral Code Linkage to Expand the NIH DS-Connect® Registry and Assess Down Syndrome Participants’ Self-Determination**

Lemuel R. Waitman, Ph.D\(^1\), Sravani Chandaka MS\(^2\), Daniel W. Connolly BS\(^2\), Jud Rhode BS\(^3\), Debbie Jae MS\(^3\), Evan Dean, Ph.D\(^4\).

\(^1\)Department of Health Management and Informatics, University of Missouri, Columbia, MO; \(^2\)Division of Medical Informatics, Department of Internal Medicine, University of Kansas Medical Center, Kansas City, KS; \(^3\)Invate Corporation, San Francisco, CA; \(^4\)Kansas University Center on Developmental Disabilities, University of Kansas, Lawrence, KS.

**Introduction**

Online cohort studies may seek to build upon national registries but are challenged when studies also collect patient reported data via validated survey websites or use existing EHR data repositories. The DS-DETERMINED study supports and enhances Down Syndrome research and the DS-Connect\(^®\) registry by leveraging PCORnet\(^1\), the National Patient-Centered Clinical Research Network (www.pcornet.org). This study links PCORnet to the DS-Connect\(^®\) (dsconnect.nih.gov) and tests capability to: 1) increase DS-Connect\(^®\) enrollment, 2) extract of clinical observations, treatments, and outcomes from PCORnet patients, and 3) conducting cognitive assessment of self-determination in the PCORnet Down Syndrome population.

**Methods**

DS-DETERMINED builds upon the PCORnet ADAPTABLE\(^2\) study’s use of a trial invitation identifier to link the de-identified pat_id identifiers used in the PCORnet Common Data Model (CDM) to the invited patient’s specific enrollment in DS-Connect\(^®\). REDCap is used to generate random identifiers centrally for each recruiting site who will track their relationship to the random identifiers used to uniquely identify patients in their site PCORnet CDM (e.g. pat_id). Site personnel then distribute these trial_invite_codes to the subject via REDCap. Upon enrollment to the DS-DETERMINED study, the subject is sent into DS-Connect\(^®\) directly using a specific study_id referral code. After enrolling and consenting to DS-Connect\(^®\), participants enter additional patient reported outcomes and upon completion, the participant is directly sent to a third website hosting the Self-Determination Inventory (SDI https://selfdetermination.ku.edu/homepage/assessments/) passing the referral code from DS-Connect\(^®\). The definitions for this specific study are logged in the PCORNET_TRIAL table. Information about completion status of DS-Connect\(^®\) surveys and SDI are sent back to the DS-DETERMINED study to ascertain how subjects are progressing so that coordinators can provide support and generate an electronic gift card upon completion.

**Results:** This linkage methodology and study framework is implemented (creating new APIs for DS-Connect and he Self-Determination Inventory) and is recruiting participants to the study.

“The study was supported by the KIDDRC NIH U54 HD 090216 and the authors acknowledge the contribution of DS-Connect\(^®\) (The Down Syndrome Registry) which is supported by the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD), NIH for the study. used in this presentation.”

**References**


A Responsive Data Curation and Integration Pipeline for Rare Disease Data

Walls, Ramona L, Ph.D.1,2, Jeffery S. Barrett, Ph.D., F.C.P.1, Alexandre Betoune, Ph.D. 1, Emily Hartley, B.S.1, Jane Larkindale, D.Phil. 1, William T. Roddy, B.S.1, Amanda Borens, M.S. 1

1Critical Path Institute, Tucson, AZ, USA; 2University of Arizona, Tucson, AZ, USA

Introduction: Rare diseases are defined in the United States as conditions or disorders that affect fewer than 200,000 people. There are over 7,000 known rare diseases, with an estimated 350 million people affected worldwide. Development of therapies for rare diseases is slow, due to limited biological understanding, challenges in running clinical trials in small, heterogenous, dispersed populations, and limited data sharing. Existing rare disease data are fragmented and often siloed. Making them interoperable is crucial but remains challenging due to lack of standardization. Recognizing these challenges, the Food and Drug Administration (FDA) initiated the Rare Disease Cures Accelerator (RDCA) that includes the development of a data and analytics platform (DAP). RDCA-DAP is a partnership between Critical Path Institute, National Organization for Rare Disorders, and others to improve the quality and the accessibility of existing rare disease data by using FAIR (findable, accessible, interoperable, and reusable) best practices, including ontologies. The platform includes sophisticated analytics (e.g., cohort building, statistical analysis, R, virtual machines) in support of development efforts and regulatory approval. Our unique approach to responsive curation bridges the gap between generalist and specialist data repositories by providing detailed curation for specific use cases with rapid discoverability of diverse datasets and disease presentations.

Data, Governance, and Privacy: RDCA-DAP currently houses over 50 rare disease datasets, including clinical trials, natural history studies, patient-reported registries, and electronic health records. All data are shared through a data contribution agreement (DCA) specifying re-use requirements and ensuring that data were gathered in an ethical manner and that individuals have provided appropriate consent. Custodians agree to share summary level data with the community, so that users may explore what data are available through the platform. Only de-identified data are included, and all data are encrypted in transfer and at rest. Users must sign a data use agreement that legally binds them to maintain the data as secure as the platform and prevents further sharing without permission.

Data Curation and Processing Pipeline: All data processing takes place on a secure cloud platform built on the Aridhi Digital Research Environment. Once a DCA is in place, the custodian shares a copy of the de-identified dataset (including metadata, data dictionaries, and supplemental documents) with the platform through a secure mechanism. A data receipt is automatically sent to the contributor to ensure correct content, including a report with feedback related to data quality improvement. A read-only copy of the data is moved to a secure data warehouse, and another copy is moved to an internal workspace for curation. All data receive a minimal level of curation that includes semi-automated generation of data dictionaries and structured metadata, and data are immediately searchable by disease and study type, variables included, and number of patients. As users discover datasets they want to analyze, a responsive curation process triggers full curation of needed variables. High priority data needed for near-term analysis are curated at the variable level by standardizing to the OMOP Common Data Model, cleaning data values (e.g., tagging outliers), and mapping to ontologies. Fully curated data are stored in both relational and graph databases in a data lake along with minimally curated data as files. Data at all levels are discoverable through a search index.

Innovation and next steps: RDCA-DAP is a neutral platform available to the rare disease community (pharma and biotechnology companies, academics, nonprofit groups, others) that provides integrated patient-level data and facilitates their analysis. The data pipeline builds off existing technologies but combines them in a novel way. It allows for scalable, standards- and ontology-based integration of diverse data sources and automated data status tracking in a responsive workflow that prioritizes curation based on need. Because fully curated data are mapped to a format-agnostic semantic knowledge graph, multiple datasets can be integrated and used together, both within a single disease area and across diseases. Future development will allow users able to export the data in multiple formats, including tabular data for analysis and CDISC-SDTM (Standard Data Tabulation Model) for regulatory uses. The platform launches in September 2021 with data from at least 20 disease areas and basic analytic features, with more to be added over time. Critical Path’s quantitative medicine and regulatory science experts can partner with groups working on common problems in the precompetitive space to develop tools and quantitative solutions.

1. https://www.govinfo.gov/content/pkg/STATUTE-96/pdf/STATUTE-96-Pg2049.pdf
2. https://fair.dap.c-path.org
The Representativeness of “People Also Ask” of Google Web Search on the Information Needs Concerning Alzheimer’s Disease and Related Dementias

Lu Wang, M.S.1, Jina Huh-Yoo, MHCI, Ph.D.1
1Drexel University, Philadelphia, PA, USA

Introduction
While caring for patients with Alzheimer’s disease and related dementias (ADRD) requires specialized knowledge and skills, the training and education caregivers receive are often not systematic or organized. To better support millions of people with ADRD and their caregivers, their information needs have been investigated by many researchers to guide the assistive technology design. Researchers applied multiple research methods including interviews, diary studies, focus groups, and questionnaires. However, several problems make it challenging to gain a whole picture of the information needs: (1) The results are generally limited to small samples due to the high resources required in recruiting participants and analyzing qualitative data; (2) It is challenging to relate the results of different studies based on the incommensurable data derived through different methodologies; (3) The results might carry bias and errors for the retrospective, self-reported studies. Systematic reviews of consolidating existing studies can be helpful, but they are costly in terms of time and effort. A more economical solution could be to collect the information needs using behavior statistics. Google Web Search, as the most widely used search engine in the United States, introduced the “People also ask” (PAA) section, which applied machine learning models, RankBrain and BERT, to identify search patterns, predict individual searches, and offer suggestions according to a search query. We compare the results of systematic reviews and that of PAA to investigate how much PAA represents the review results.

Method
Information needs summarized in systematic review and scoring review in the past three years introduce four themes with thirty-eight specific information needs from 50 original studies in total1-3. We collected the data by entering the terms “dementia” and “alzheimer’s” into Google Web Search (www.google.com) and generating the list of frequently associated questions from PAA on the search result page. Clicking on each question on PAA will populate 1-2 more questions. For each search term result, we clicked 100 questions in its ranked order. As a result, we collected 574 questions in total (282 for “dementia” and 292 for “alzheimer’s”) and classified the questions according to the references. A clean-installed Chrome browser was applied to avoid the influence of personalized recommendations. We used the Scrapper (v.1.7) tool to download the questions. The date for this collection was January 4th, 2021.

Results
Results showed that, for the theme of “disease-specific information,” questions generated from PAA covered 91.67% of the 12 information needs identified in1-2, but missed the information need of “current research on ADRD.” For the theme of “patient care provision information”, questions generated from PAA covered 73.33% of the 15 information needs, but missed “safety issues like how to improve safety of environment, how to keep patient safe, how to recognize fall risks and poor mobility”, “emergency situations”, “helpful experiences of other caregivers” and “how to deal with family and friends.” For the theme of “healthcare service-related information”, only 14.29% of the 7 information needs were covered, missing “where and how to use services/help available”, “legal issues”, “how to apply for care programs eg. daycare, long term care”, “insurance”, “home help” and “transportation options.” For the theme of “caregiver self-care,” none of the 4 information needs was covered, including “stress management”, “caregiver’s entitlements (pension)”, “managing emotions” and “general caregiver self-care activities”.

Conclusion
The representativeness of PAA varied among the themes of information needs, possibly resulting from different usage habits on different platforms. PAA can be an efficient tool to identify specific areas of information needs. Future work should investigate the patterns of which information needs PAA represents better than others. Furthermore, we can explore combining behavior statistics from multiple aggregate search behavior tools (e.g., Google Trends, Alexa).

References
Factors Affecting Telemedicine Utilization among Cardiovascular Disease Patients During the COVID-19 Pandemic
Suwei Wang, PhD1, Hu Huang, PhD1, Elisabeth Lee Scheufele, MD, MS1, Gretchen P. Jackson, MD, PhD1,2, Irene Dankwa-Mullan, MD, MPH1
1IBM Watson Health, Cambridge, MA; 2Vanderbilt University Medical Center, Nashville, TN

Introduction
The COVID-19 global pandemic severely disrupted patterns of in-person outpatient visits and contributed to delayed or limited care of patients with chronic diseases. Telemedicine expanded rapidly as a safe and viable means to extend access to routine care while reducing COVID-19 exposure risk for vulnerable patients and health professionals. Early evidence pointed to higher risk of disease and poorer outcomes for patients with cardiovascular disease (CVD). In this cohort of commercially insured individuals, we sought to examine demographic and patient-level factors associated with patients with CVD who attended telemedicine visit vs who didn’t (“no-shows”).

Methods
This retrospective study examined adult patients (age ≥18 years) with CVD and telemedicine appointments using the IBM® Explorys® Database, a data set of United States (US)-based electronic medical records, using Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) codes for CVD and internally developed code-sets, in the timeframe of 03/01/2020 to 09/30/2020. Telemedicine encounters were identified through eligible Current Procedural Terminology (CPT®) codes. Patient demographic (e.g., sex, race, ethnicity, insurance) and clinical characteristics (e.g., liver disease, renal disease) were identified and analyzed. Univariate and multivariate logistic regression were performed to identify patient-specific factors associated with a completed telemedicine visit.

Results
A total of 744,485 patients with CVD were identified with 1,548,368 scheduled telemedicine appointments, of which 376,118 patients (50.5%) attended their appointment. There were 714,965 completed telemedicine encounters (some patients had more than one completed telemedicine visit). Mean age of patients who attended at least one telemedicine visit was 59 years (range 18-90 years), with demographic distributions: 60.5% female, 74.0% Caucasian, 14.5% African American, 1.2% Asian, 5.3% Hispanic, 1.5% Other, 98.0% English-speaking, 10.3% self-pay, 50.4% commercial insurance (private), and 29.0% on Medicare. Multivariate analysis identified the factors associated with lower odds of attending telemedicine appointment including male sex (OR=0.75, 95%; CI:0.75-0.79), African American race (OR=0.81, 95%; CI:0.76-0.87), speaks languages other than English or Spanish (OR=0.43, 95%; CI:0.36-0.54) and self-pay (OR=0.78, 95%; CI:0.73-0.85). Patients with comorbid chronic liver disease (OR=1.32, 95%; CI:1.19-1.47), kidney disease (OR=1.27, 95%; CI:1.16-1.40), hyperlipidemia (OR=1.12, 95%; CI:1.20-1.35), cancer (OR=1.25, 95%; CI:1.18-1.31) and with Medicare (OR=1.11, 95%; CI:1.03-1.19) had higher likelihood of completing telemedicine appointment.

Discussion
In this study during the early COVID-19 pandemic, only half of patients with scheduled telemedicine appointments showed up to their visit. Factors associated with telemedicine no-shows were similar to characteristics associated with barriers to in-person care. However, medical complexity and having insurance coverage were associated with telemedicine visit completion. As such, individuals who are male, African American, speak languages other than English or Spanish, or are self-pay may require additional efforts to ensure that barriers to attending telemedicine are anticipated and addressed.

Conclusion
Telemedicine has rapidly become an essential resource for patient care during the COVID-19 pandemic. In this study, patients with comorbidities and Medicare were more likely to attend telemedicine appointments, and those patients who were male, African American, speaking languages other than English or Spanish, or self-pay were less likely to complete telemedicine encounters. This study has identified gaps in equitable telemedicine access. Additional research is needed to determine whether these trends continue as telemedicine adoption increases and to understand the causes of these association to develop effective mitigation strategies.
Introduction

Chest X-ray images have proven useful for monitoring various types of the disease, including COVID-19 and related disease diagnosis. The rapid spread of COVID-19 has caused a global pandemic with variants Delta and Beta spreading fast. Deep learning based models for COVID-19 detection has shown high performance. However, there is a gap between the current deep learning-based classification and the clinical needs: detecting subtle differences in the medical images. In addition, current supervised training of deep neural networks on medical image diagnosis relies on large pools of labeled data, which is scarce, inflexible, and limited by types of annotations. A flexible approach that can reflect user-defined criteria without a large amount of annotation is needed. Further, the approach needs to enable collaborated real-time labeling from radiologists distributed across different geographic locations. In this study, a 5G-based active learning framework is proposed for selective annotation to magnify the labeling effect, precisely connecting the medical images to corresponding patients’ history records, support secure distributed annotation across multiple locations and enables low latency near-real-time decision support.

Methods

Existing CNN-based supervised deep learning CheXNet shows high accuracy in 14 common diagnoses in Chest X-rays. Grad-CAM\textsuperscript{1} introduced for ‘visual explanations’ for decisions. Based on the two models used the last layer of Full connection and the last layer of deployed CheXNet models as Feature construction for active learning and initial domain knowledge injection. Instead of identifying pre-defined diagnoses, the proposed system enables a user-defined target through the labeling of images. Such labeling has multiple sources: radiologist and medical experts, cross-sourcing patients’ history, EHR records or referencing other models, etc.

An end-to-end 5G network slice with Multi-access Edge Computing (MEC) is being built to enable a low latency, secure communication of medical images, and cross-platform patients’ medical records, distributed annotation, and near-real-time decision support. Through active learning, labeling is not a prerequisite but an ongoing process with utilizing the model. The key for active learning is selecting the right sample for interactive labeling. The specified domain knowledge is introduced through label propagation, enables user-defined target, maximizes the value of labeling through active learning by introducing the concept of records resolution. The labeled image improves model performance by guiding the direction of the model.

Results and Future Work

The proof of concept model accuracy was evaluated by the average value of F1 scores at different sample percentages. As the percentage of the labeled sample increased, the F1 score also increased. At the percentage of 40\%, F1 score ranged from 0.86 to 0.96 in between different diagnoses\textsuperscript{2}. Further evaluation of the model is still in progress. The Future work includes: incorporating unlabeled samples into training through Generative Adversarial Networks (GAN), Cross-referencing can directionally search the patient’s historical records related to the target definition and incorporate policy, regulation, and labeling costs into the system.

References


Introduction

As the world searches for effective treatments and potential cures for the COVID-19 pandemic, the ability to consolidate data, insights, and expertise from many disparate sources will be key to fully understanding the patient outcomes of the infection. At Mass General Brigham, we have developed a cohesive research data portal called the COVID Biobank Portal that enables clinicians, researchers, data scientists, and technologists to work together with a common goal of better understanding COVID-19 symptoms, associated risk factors, and successful therapies. The novel data portal brings data, analytics, and applications together in a way that provides immediate value for the current COVID-19 crisis while also unlocking the bench-to-bedside capabilities that researchers currently need to advance research.

Methods & Discussion

At Mass General Brigham, the COVID Biobank Portal is an Integrating Informatics Bench to Bedside (i2b2) instance that contains Electronic Health Record (EHR) data for over 33,000 COVID-tested patients already consented within the Mass General Brigham Biobank. Genomic array and exome sequence data is also available for a subset of these patients. Within the data portal, EHR data types are classified into an ontology system that groups medical terms and links directly to the underlying data—these groupings include encounter variables, demographics, vital signs, personal health and family history survey data, diagnoses, procedures, medications, laboratory results, blood-sample availability, and, where available, genomic data.

The COVID Biobank Portal is not only used to query for patients, but also provides the ability for users to download data files of limited data sets (LDS) for further cohort analysis. The downloaded file format includes de-identified patients in a customizable single-patient-per-row file with curated data in the columns. The novel approach to customizing the patient-level data file is key to producing a file with temporal constraints related to COVID-19. For instance, we have provided the ability for the user to designate one column as the index date. You can define an offset-in-days around that index date, for example, whether a patient was on a medication within one year to one week before they were tested and found COVID positive.

Results & Conclusion

The COVID Biobank Portal has been implemented and released to a group of beta testers made up of members from the ‘Using EHR data for COVID-19 research’ Working Group, an initiative of the Mass General Brigham Center for COVID Innovation. The operation and maintenance of the novel data portal is considered ‘production-ready’ in which data is refreshed and new COVID-tested consented patients are added on a weekly basis. The initial response from the beta testers has been overwhelmingly positive and the utility of the data portal has been demonstrated.

Building a focused data portal for COVID-19 can help facilitate analytics and knowledge management of the COVID-19 pandemic. The development of novel tools that work with the i2b2 platform allowed for the creation of the COVID Biobank Portal to be flexible to disparate data sources and highly scalable in producing useful data files for further COVID-19 research.

References

The PROVEN Coordinating Hub to Accelerate Research about Electronic Health Record Modernization in the U.S. Department of Veterans Affairs

Michael Weiner, MD, MPH1,2,3, Elizabeth M. Yano, PhD, MSPH4,5,6, Jessica A. Davila, PhD, MS7, Catherine Brayton, MPH4, Alison M. Cogan, PhD4, Seppo Rinne, MD, PhD8, Steven R. Simon, MD, MPH4,6

1Center for Health Information and Communication, U.S. Department of Veterans Affairs, Veterans Health Administration, Health Services Research and Development Service CIN 13-416, Richard L. Roudebush VA Medical Center, 2Regenstrief Institute, 3Indiana Univ., Indianapolis, IN; 4VA HSR&D Center for the Study of Healthcare Innovation, Implementation & Policy, VA Greater Los Angeles Healthcare System, 5Department of Health Policy & Management, UCLA Fielding School of Public Health, 6Department of Medicine, UCLA Geffen School of Medicine, Los Angeles, CA; 7Michael E. DeBakey VA Medical Center, Houston, TX; 8VA Bedford Healthcare System, Bedford, MA; U.S.A.

Introduction

The world’s largest transition between electronic health record (EHR) systems is occurring in a $16 billion implementation in the U.S. Department of Veterans Affairs (VA). Research is needed in clinical outcomes, human factors, decision support, visualization, methodology, data access, training, satisfaction, costs, and the advancement of quality, safety, and value. To foster research in these areas, the VA Health Services Research and Development Service has funded a Coordinating Hub to Promote Research Optimizing Veteran-Centric EHR Networks (PROVEN).

Aims of the Coordinating Hub

PROVEN’s aims are to support, promote, & enhance health services research related to EHR modernization; provide methodological training, methodological and informatics expertise, and support to investigators; serve as a liaison between EHR modernization researchers and VA operational partners and implementation sites; communicate research findings to VA decision makers and the research community; and oversee the conduct of pilots investigating EHR implementation.

Organizational Overview

PROVEN has four cores: administrative; health informatics; data and methods; and implementation & training. PROVEN regularly interacts with stakeholders including the VA Office of EHR Modernization. An initial set of core pilot projects have been developed, covering areas of governance, clinical care, and Veteran engagement.

Progress to date

The PROVEN Hub has held online seminars, established an investigative network, and developed and disseminated a research agenda for EHR modernization. Projects in data quality and data training are in development. Discussions with partners led to expanded involvement of researchers in EHR Councils and site-level planning for implementation. Enhancing enterprise-wide access to EHR data for research remains a priority. Published articles on EHR transitions have been identified. Findings are being methodically mapped and analyzed, to synthesize lessons that may guide the implementation efforts.

Conclusion

The PROVEN Hub has created necessary research infrastructure to support and foster research about EHR modernization in VA. Pilots have been initiated, but the bulk of the critical work lies ahead and includes not only research and implementation science but also attention to workforce capacity in informatics research.

Acknowledgements

Source of funding: VA HSR&D SDR 20-197. Views expressed herein are those of the authors and do not necessarily represent the views of the U.S. Department of Veterans Affairs.
Nursing Data in a CoVID-19 Vaccination Clinic

Luann Whittenburg, PhD, RN BC, FNP BC, FHIMSS, FAAN, FAMIA1
Virginia K. Saba, EdD, RN, FACMI, FAAN LL2
1Health Informatics Consultant, Fairfax, VA
2President and CEO, SabaCare, Inc. Arlington, VA

Keywords: Data Sharing, Interoperability, Knowledge Representation, Nursing Informatics, Clinical Care Classification System, CCC

Overview: The Coronavirus 2019 (CoVID 19) pandemic is a global health crisis. Frontline nurses have suffered psychological stress and prior studies have focused on knowledge of the disease and information sources for the public.1-2 In January 2021, CoVID-19 vaccination clinics opened, and a literature review identified a data gap. No standard, coded nursing data was associated with the CoVID-19 vaccination process. For example: nursing services required to discuss potential vaccine side effects, post-vaccine care, injection administration or nursing services for infection control. This project identifies the relationship between nursing concepts in a recognized nursing informatics standard and the nursing services provided in a CoVID-19 vaccination clinic to evaluate if nursing services data from CoVID-19 vaccination clinics could be coded and if so, can standardized, coded nursing data enhance the granularity of nursing data pertaining in CoVID-19 vaccine administration predictive workload models. The recognized nursing terminology of choice for documenting the essence of patient care in electronic health record systems is the Clinical Care Classification System (CCC) classified by 21 care components to form a single interrelated system that provides a framework linking nursing diagnoses to interventions, and outcomes as well as linking to other health-related terminologies.3-4 The CCC four level framework allows aggregation of coded data upward for big data analytics and downward to granular data. The CCC has a five-character alpha-numeric structure for data interoperability based on the ICD-10.5

2. Objective: The objective was to code the nursing services performed during a CoVID-19 vaccination clinic and identify the actual nursing services involved in the vaccination process in a CoVID-19 vaccination clinic.

3. Methodology: The design was a deductive content analysis based on degrees of equivalence often used to match concepts between defined process documentation and coding systems. Of the available 201 CCC Nursing Interventions and four CCC action types the CCC nursing intervention concepts total of 804 concepts. The CCC concepts were input into an Excel worksheet matched to CoVID vaccine documentation based on degrees of equivalence. Matches were reviewed by CCC content experts with differences discussed until 100% consensus agreement was reached. With agreement reached, the match outcome was entered on the Excel abstraction form.

4. Result: All 10 of the unique nursing services in the CoVID-19 vaccination documentation matched a CCC concepts. The CCC demonstrated an ability to code nursing services at the smallest data level and the content analysis process data enhanced CoVID vaccination documentation. The CoVID nursing services data matched to CCC concepts provided a cohesive, objective view of the CoVID vaccination process and realistic assessment of the nursing services provided in a CoVID-19 vaccination clinic.

5. Discussion: The CCC System is designed for computer processing, free for use with copyright permission, approved by the ANA, HHS, and conforms to Cimino criteria6. The CCC increased the descriptive validity of the CoVID-19 vaccination documentation and data consistency between CoVID vaccine documentation and CCC nursing concepts. Future studies may access the CCC Workload Actions Measures Method6 (WAMM)7 to generate valid, reliable measurement of nursing workload through weighted CCC ‘care value’ (intervention/action types) coupled with CCC ‘base values’ (health conditions acuity).

6. Conclusion: Nursing documentation in a public health system using coded nursing concepts in the CCC System is important to understanding the actual workload of nursing practice in a CoVID-19 vaccination clinic. This CCC application demonstrates nursing services data from clinic vaccination processes can be coded in a recognized nursing informatics standard and nursing data increases the granularity of available nursing data.

7. References
3. Healthcare Corporation of America (HCA) https://careclassification.org/about/
Complementing Automated Risk Prediction with Face-to-face Screening Improves Suicide Risk Prediction

Drew Wilimitsis1, Robert W. Turer, MD1, Michael Ripperger1, Allison B. McCoy, PhD1, Sarah H. Sperry, PhD1, Colin G. Walsh, MD, MA1; 1Vanderbilt University Medical Center (Nashville, TN, United States)

Background
In a systematic review, the U.S. Preventive Services Task Force found insufficient evidence to support the accuracy and effectiveness of suicide risk screening in primary care settings. Although machine learning can leverage electronic health record (EHR) data for scalable risk detection, suicide prediction models have high false-positive rates that limit their clinical utility. Our team at Vanderbilt University Medical Center (VUMC) has developed and implemented a real-time suicide risk prediction model across the enterprise. To investigate the respective and combined predictive abilities of in-person screening and an automated prediction model, we evaluated (1) the ability of the Columbia-Suicide Severity Rating Scale (C-SSRS) to predict suicide attempt and ideation, (2) compared the performance of the C-SSRS to our risk model, and (3) analyzed the performance of our risk model combined with the C-SSRS to evaluate synergistic effects.

Methods
We studied an observational cohort of all adult patients from June 2019 to September 2020 that received both in-person screening and real-time risk predictions. This cohort included visits in the inpatient outpatient, and emergency department (ED) settings. At each index visit, we extracted C-SSRS responses and VSAIL (Vanderbilt Suicide Attempt & Ideation Likelihood) scores generated by our prediction model. To assess the predictive validity of the C-SSRS, we trained a logistic regression model using binary features that represented patient screening responses. Then, we compared the performance of the C-SSRS, VSAIL, and ensemble models (Lasso regression, unweighted average, and a weighted average based on screening risk) that combined both predictions.

Results
After applying inclusion criteria, 120,398 unique patient encounters were included in the dataset. Logistic regression odds ratios for significant C-SSRS predictors were Q1 (SA: 8.8, SI: 10.59), Q2 (SI: 2.04), Q5 (SA: 2.35), and Q6 (SA: 2.46, SI: 1.56). Within thirty days of an index visit, combined models had higher AUC (SA: 0.874 - 0.887, SI: 0.869-0.879) than both the VSAIL (SA: 0.729, SI: 0.773) and C-SSRS (SA: 0.823, SI: 0.777). In the top 5% of risk scores, ensemble methods had PPV (SA: 2.6%, SI: 15.2%) and sensitivity (SA: 76.1%, SI: 61.4%), significantly improving over the VSAIL model, which had PPV (SA: 0.6%, SI: 5.2%) and sensitivity (SA: 16.1%, SI: 20.9%).

Conclusion
Ensemble models combining C-SSRS responses and an EHR-based model performed significantly better than either alone in the prediction of suicide attempt and ideation at a variety of time intervals. By incorporating patient screening responses, the PPV for suicide attempt in the highest risk ventile surpassed an acceptable threshold required for effective clinical decision support. Our results suggest that traditional clinical assessment and predictive models can be used in complement to overcome their respective limitations, which currently hinder the ability of suicide risk detection systems to meaningfully inform clinical practice.

References
Classifying Infection Risk Following Pediatric Cardiac Surgery

Kaitlin C. Williamson, M.D.1, Daniel Fabbri, Ph.D.1
1Vanderbilt University, Nashville, TN

Introduction
Each year in the United States, approximately 40,000 children undergo cardiovascular surgery, with infections complicating as many as 30%.1,2 If patients at higher risk of infection could be identified prospectively, clinical decision support could prompt increased infection prevention measures and improve outcomes. A past study in this population estimated risk of post-operative septicemia, mediastinitis, and endocarditis, with fairly good performance (c index 0.79), but this study missed many clinically relevant infections, and used only regression methods.1 Our aims in this study were to generate a cohort of pediatric cardiac surgery patients, to generate models to classify the risk of infections in the 30 day postoperative period, and to compare the results of regression and machine learning methods.

Methods
We obtained our data via query of the Vanderbilt Research Derivative, a clinical data warehouse repurposed for research. We included any pediatric cardiac surgery performed on or after January 1, 2015, on patients under age 18. We included 36 predictors including demographics, lab values, comorbid medical conditions, and drug exposures. Outcomes included sepsis, bacteremia, urinary tract infection (UTI), surgical site infection (SSI), pneumonia (PNA), necrotizing enterocolitis (NEC), and a composite outcome of any previously listed outcome excluding NEC, occurring within 30 days of the procedure. We used Python’s scikit-learn package for analysis. We performed an 80-20 test-train split, then fit and ten-fold cross validated models using logistic regression, support vector machine (SVM), K-nearest neighbors (KNN), decision tree and random forest models. We assessed performance using area under the receiver operating characteristic curve (AUROC) and accuracy. We performed univariate analysis of predictors for each outcome and performed subgroup analyses of younger (<180 days) and older children (>=180 days).

Results
We obtained data for 1681 patients, undergoing 2080 distinct operative encounters. Our univariate analysis of the full data set, including all ages, showed that patient age was the strongest predictor for all outcomes. Table 1 describes the performance of each model in the test data, in classifying each outcome.

<table>
<thead>
<tr>
<th></th>
<th>Sepsis (n=75)</th>
<th>Bacteremia (n=114)</th>
<th>UTI (n=57)</th>
<th>SSI (n=89)</th>
<th>PNA (n=89)</th>
<th>NEC (n=40)</th>
<th>Composite (n=357)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic Regression</td>
<td>0.921 (98%)</td>
<td>0.781 (94%)</td>
<td>0.701</td>
<td>0.694</td>
<td>0.708</td>
<td>0.946</td>
<td>0.759</td>
</tr>
<tr>
<td>SVM</td>
<td>0.772 (98%)</td>
<td>0.708 (94%)</td>
<td>0.716</td>
<td>0.622</td>
<td>0.708</td>
<td>0.942</td>
<td>0.659</td>
</tr>
<tr>
<td>KNN</td>
<td>0.674 (97%)</td>
<td>0.674 (94%)</td>
<td>0.568</td>
<td>0.535</td>
<td>0.568</td>
<td>0.812</td>
<td>0.666</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>0.614 (97%)</td>
<td>0.541 (94%)</td>
<td>0.523</td>
<td>0.538</td>
<td>0.528</td>
<td>0.684</td>
<td>0.636</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.606 (97%)</td>
<td>0.672 (94%)</td>
<td>0.751</td>
<td>0.623</td>
<td>0.751</td>
<td>0.803</td>
<td>0.604</td>
</tr>
</tbody>
</table>

Table 1. AUROC in test set for full cohort (n = 222).

Discussion
Several of our models performed well in classifying post-operative infection risk. Compared to machine learning methods, logistic regression performed more consistently between the training and test data, with particularly high performance in classifying NEC, bacteremia, and sepsis. Several models demonstrated likely overfitting. If validated in real time data and applied clinically, these models could provide useful forecasts of infection risk.

References
Exploring EHR Diagnosis Recording Pattern Changes During a Pandemic

Kala S. Wilson, MPA,1 Brian J. Wells, M.D. Ph.D.,2 Jyoti Gangur, MS,1
Kristin M. Lenoir, MPH,1 Maria E. Mayorga,3 Franck Diaz-Garelli, Ph.D.1
1University of North Carolina at Charlotte, Charlotte, NC
2Wake Forest School of Medicine, Winston Salem, NC
3North Carolina State University, Raleigh, NC

Introduction
Improving learning health systems rests on reliable secondary use of Electronic Health Record (EHR) data;1 however, the consistency and validity of secondary analyses are directly reliant on the data’s quality (DQ).2 We hypothesized that an external societal change (e.g., COVID-19 and increased telemedicine) may be correlated with clinical data entry patterns. We tested this hypothesis on diabetes-related DX data given its chronic, varying nature over time (i.e., controlled vs. uncontrolled levels, complex comorbidities) and significance to public health. Specifically, we built statistical models predicting the odds of DX sequences occurring before or after an elevated hemoglobin A1c (HbA1c) based on the number of DX records that indicated uncontrolled diabetes to compare effects during regular and confinement operations.

Methods
We assessed changes in reporting rates of uncontrolled diabetes in structured diagnosis (DX) data preceding and subsequent to a lab indicating this transient condition was available in the EHR. We compared patterns during normal clinical operations and during the current pandemic’s Stay at Home orders of 2020 in North Carolina. After receiving approval from the Wake Forest University School of Medicine’s Institutional Review Board (IRB: IRB00062976), we extracted structured DX data from an EHR database covering five segments within electronic patient charts (i.e., primary encounter DX, encounter DX, billing DX, order DX, and problem list DX). We gathered patient DX sequences (i.e., a chain of chronologically ordered DX records in a patient’s EHR) including all diabetes DXs within 90 days before and after a HbA1c laboratory result greater than 9%.3 Every DX sequence contained specific ICD-10 codes for Type 1 or 2 diabetes (i.e., ICD-10 DX code, E10.* and E11.*) that indicated lack of control (i.e., “uncontrolled diabetes”, “diabetes with hyperglycemia”). We tested our hypothesis by building binomial statistical models using Stata’s generalized linear model (GLM).

Results
A total of 107,167 diabetes DX recordings for 12,620 patients and 6,574 diabetes DX recordings for 1,187 patients for regular operations and pandemic operations respectively were assessed. Our statistical models revealed changes in number of DX reporting uncontrolled diabetes between regular clinical operations and clinical operations during confinement. A 90-day DX sequence was 6.9% (adj-p<.0001) more likely to appear after a high HbA1c value during regular operations for each uncontrolled diabetes DX included. In contrast, it was 15.9% (adj-p=.0094) more likely to appear in the post-HbA1c timeframe during confinement operations. We also found statistically significant differences for number of providers and number of departments predicting whether a DX would appear before or after the elevated HbA1c between normal (OR=1.309; OR=2.405, adj-p<.0001) and confinement (OR=1.722, adj-p=.0101; OR=4.107 adj-p<.0001) clinical operations. After controlling for timeframe (i.e., normal operations and confinement operations), we found the effects were statistically different (adj-p<.0001) with an odds ratio of 1.20, revealing higher odds of uncontrolled diabetes DX appearance in the normal operations dataset.

Conclusion
We were able to uncover differences between regular and confinement operations, providing preliminary evidence to the idea that pandemic-induced alterations of clinical workflows may be associated with clinical data quality and entry patterns. Our findings emphasize the need to give particular attention to factors impacting clinical practice while generating secondary use of EHR data. Future work will verify the internal and external validity of these findings and further investigations will explore the causal relationship between data entry changes and the onset of external factors.

References
Prediction of Small Bowel Obstruction Outcomes with Machine Learning

Tanner Wilson, MD, MS1,2, Gilles Clermont, MD, MSc1,2, Harry Hochheiser, PhD2
1UPMC, Pittsburgh, PA; 2University of Pittsburgh, Pittsburgh, PA

Introduction
Small bowel obstructions (SBOs) are a common cause of hospital admissions for abdominal pain. Surgery is required for 24% of cases, and when indicated, early operation reduces mortality by half1. Unfortunately, clear indications for early operation are elusive. For example, CT sensitivity for intestinal ischemia, a clear indication, is as low as 75%2. One influential model of SBO3 based on symptoms and CT findings was 37% sensitive for need for operation and 46% sensitive for strangulated obstruction, with accuracy of 65% and 77%. We aim to leverage large-scale informatics and machine learning methods to improve the prediction of outcomes in SBO by creating a screening model using structured data available within the first day of admission to be deployed on all admissions with an SBO-like diagnosis.

Methods and Results
We collected electronic medical record extracts from 28,298 admissions between 2010-2019 across 15 UPMC hospitals with ICD9/10 coding consistent with small bowel obstruction of any cause, as well as possible mimics such as ileus, at any point during admission. From these records, day-of-admission observations were collected on the following: age, sex, race, BMI, medical comorbidities as defined by Elixhauser4, smoking status, and basic laboratory values (hematologic, electrolyte, renal/hepatic function, and coagulation panels). Any ICD9/10 code noting hernia as the cause of intestinal obstruction was also noted, as this was thought to influence outcomes. Chief complaints were reviewed to assure an appropriate cohort, with the top 5 explicitly specifying SBO or abdominal pain (not shown).

All computations were performed in Python 3.7.3 using scikit-learn 0.24.1. Admissions beginning in 2018 or later were treated as test cases, for 18% holdout. Numerical variables underwent mean imputation and standard scaling.

Outcomes of interest included: any relevant abdominal operation during the admission (50% of cases), as identified by ICD9/10 coding; any intestinal resection during the admission (23%), by ICD9/10; and patient death during the admission (5.7%), by date of death recorded in the EMR. For each outcome, we trained models using logistic regression with $L_2$ regularization, random forests, k-nearest neighbors with both uniform and distance-based weighting, support-vector machine classification using both linear and radial basis function kernels, and Gaussian naïve Bayes classification. Hyperparameter optimization was performed with 3-fold cross-validation to optimize the F1 metric. Balanced class weighting was used for all applicable models to account for class imbalances. For each outcome, models with at least 50% recall of positive cases were then combined into hard and soft voting ensemble classifiers, with the final voting method chosen for highest recall.

Conclusion
Our initial model achieved recall of 72% of operative cases, 57% of bowel resections, and 55% of inpatient deaths. As seen in the table, overall performance is promising for an initial screening model based exclusively on structured data, especially for the highly imbalanced inpatient deaths. This represents the first step in development of a decision support system for SBO at our institution. Future steps will expand to additional sources of both structured and unstructured data, including the text of radiology reports and clinical notes, to create more clinically informed models.

<table>
<thead>
<tr>
<th>Test Set Results</th>
<th>Operation</th>
<th>Voting</th>
<th>F1</th>
<th>Recall</th>
<th>Precision</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Soft</td>
<td>0.62</td>
<td>0.72</td>
<td>0.54</td>
<td>0.57</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Soft</td>
<td>0.27</td>
<td>0.57</td>
<td>0.18</td>
<td>0.56</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Hard</td>
<td>0.28</td>
<td>0.55</td>
<td>0.19</td>
<td>0.82</td>
<td></td>
</tr>
</tbody>
</table>

References
The Use of Integrated Medical Devices and Clinical Decision Support in the Acute Care Setting: A Scoping Review

Jennifer B. Withall, PhD, RN1, Jessica M. Schwartz, MPhil, BSN, RN1, John Usseglio, MPH2, Kenrick D. Cato, PhD, RN1,3

1Columbia School of Nursing, New York, NY; 2Columbia University Irving Medical Center, Augustus C. Long Health Sciences Library; 3Columbia University Department of Emergency Medicine, New York, NY

Introduction
Seamless data integration between point-of-care medical devices and the electronic health record (EHR) are elemental to clinical decision support systems (CDSS). CDSS directly aids clinical decision making by matching evidence-based computerized knowledge, contextual assessments, and individual patient characteristics to generate patient-specific recommendations to the clinician1. Previous reviews examined patient-generated data through consumer wearables and integration with the EHR in the community setting. This review will focus on integrated devices that clinicians encounter in the hospital setting. The objective of this scoping review is to examine the existing evidence pertaining to integrated medical devices and how the clinical decision support systems associated with these devices impacts clinicians and clinician decision making in the inpatient acute setting.

Methods
The guidelines for scoping reviews outlined in the JBI Manual for Evidence Synthesis and the Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extensions for Scoping Review (PRISMA-ScR) served as the framework for this review2. PubMed, CINAHL, IEEE Xplore, and Scopus were searched for scholarly, peer-reviewed journals indexed between January 1, 2010, to December 31, 2020. A priori inclusion criteria were empirical studies describing integrated medical devices impacting patients at the point-of-care or integrated medical devices that facilitated clinical screening that referenced solicited or unsolicited CDSS and clinician decision making, in the inpatient acute care setting. Articles that were excluded were those not available in English, situations where medical devices were not integrated with the EHR, guidelines, systematic reviews, conference proceedings, articles that examined diagnostic or imaging studies, and consumer (commercially available) wearables for physiological data (e.g., FitBit). Articles and data extraction were organized according to the initial use scenarios described in the modified Conceptual Model for Technology, Nursing and Patient Safety. The modification included changing “nurse” in the original model to “clinician” noted in this review (Figure 1)3.

Results
Of the 1924 articles screened, 18 were included for synthesis (Figure 2). Ten articles described patient assessment, monitoring, or surveillance use, with vital signs monitoring and CDS related to early warning detection comprising most of this subset. Three articles described patient protection from harm, all of which focused on CDS for medication safety. Four articles described direct care use scenarios, all of which described insulin administration and a CDS tool for insulin infusions. Finally, one article described a hybrid situation of patient communication and monitoring.

Conclusion
Integrated device data provides insight into user-device interactions and helps to illustrate healthcare processes. Considering the emphasis on patient safety, cost-effective and efficient care delivery, our results indicate that establishing closed-loop and/or bi-directional communication related to point-of-care devices remains an area of opportunity for healthcare organizations. Additional research may include integrated device log metadata, deep learning analytics, and situational awareness; artificial intelligence and CDSS; patient portals, mobile health, and incorporation of patient-generated data into the EHR and CDSS in alternate care settings; and interprofessional innovation for new knowledge generation. These future directions can broaden and extend the existing evidence related to integrated medical devices, CDSS, and patient outcomes.

Acknowledgements
This work is supported by the NINR Reducing Health Disparities through Informatics training award T32NR007969.

References
Do Clusters and Sequence of Symptoms Predict COVID-19 Test Results?

Janusz Wojtusiak, PhD, Wejdan Bagais, BS, Hedyeh Mobahi, MS, Elina Guralnik, MS, Amira Roess, PhD, Farrokh Alemi, PhD
George Mason University, Fairfax, VA

Introduction
Despite wide availability of COVID-19 vaccines, the need for frequent screening will continue. A polymerase chain reaction (PCR) tests are considered the gold standard in terms of testing accuracy with reported recall (sensitivity) over 95%. Unfortunately, it typically takes 24-48 hours for individuals to get the PCR test results which is often perceived too long in many daily situations. In contrast, rapid (antigen) tests, including in-home tests, are less accurate with recall that varies between 49% and 80% depending on a source. Alternatively, symptoms can be used.

Methods
Data: Two independently collected sources of data have been used. The 461 data were collected by our group as part of a study to investigate the relationship between symptoms, demographics, and social distancing. The self-reported data include 23 symptoms, along with indicators which of the symptoms were present on the first day. The Mason Health™ is a symptom and exposure screening tool that was implemented by George Mason University as part of the response to COVID-19 pandemics. The data include 1,396,797 daily reports by 54,805 individuals, and 261,011 test results, transformed into symptom episodes by a time-series analysis, resulting in 703 symptom-test episodes with for 594 individuals. All data analysis has been completed using the 461 data, split into 80-20 training and testing sets, bootstrapped 30 times with average values reported. The Health™ data were used external validation only. Symptom Clustering: Three approaches have been used: (1) An exhaustive set of conjunctions of pairs and triples of symptoms have been created, and passed through 24-fold cross-validated LASSO to detect robust predictors. (2) A set of all hierarchical clusterings was extracted and passed through LASSO, as for the previous method. (3) A set of clusters created by cutting the hierarchical clustering dendrogram at a specified point was used. Supervised Learning: Logistic regression, random forest and gradient boosting were used. The predicted output is positive/negative test result for COVID-19. First, models were learned using all 23 symptoms in the 461 data. Then, a subset of 16 symptoms common among the datasets was selected. Final models were applied to the Health™ data for the external validation. Area under ROC (AUC), recall and precision averaged from the 30-time bootstrapped training/testing split were calculated. Models were checked for biases and performance within minority groups.

Results
The best result on the 461 data was obtained from logistic regression on data with first day symptoms and no clusters (AUC=0.7984; 12 symptoms). The selected best model externally validated on Health™ data achieved AUC=0.7125. Symptom clusters: Inclusion of clusters of symptoms slightly improves accuracy of the models, but at the expense of the numbers of variables and overfitting. The best model with symptom clusters reached AUC=0.7981 vs. 0.7903 (p<0.01) when using clusters of variables extracted from hierarchy and achieved AUC=0.6955 on Health™ data. First-day symptoms: The experiments also indicate that the inclusion of additional attributes that indicate which symptoms were present on the first day slightly improves model performance from AUC=0.7903 to AUC=0.7984 (p=0.02) without clustering and from AUC=0.7883 to AUC=0.8004 (p=0.02) with conjunctions of symptoms (0.7 on Health™).

Conclusion
The results indicate that it is reasonable to predict COVID-19 using self-reported symptoms. The inclusion of clusters and order of symptoms improves accuracy of the models. The constructed models tend to overfit the data because of small sample size and large number of variables, which is further confirmed by the best external performance on Health™ data achieved by the simplest model, yet still providing good results.

References
Implementing Real-Time Prescription Benefit Tools: Early Experiences at 5 Academic Centers


1Stony Brook University Hospital, Stony Brook, NY; 2 Yale University, New Haven, CT, 3Yale School of Medicine, New Haven, CT; 4Johns Hopkins University School of Medicine, Baltimore, MD; 5Froedtert and Medical College of Wisconsin, Milwaukee, WI; 6Yale New Haven Health, New Haven, CT, 7University of Pittsburgh School of Medicine, Pittsburgh, PA

Background: High drug costs in the U.S. negatively impact medication adherence and clinical outcomes1. Multiple e-prescription vendors now provide real-time prescription benefit (RTPB) tools that display patient out-of-pocket prices in the electronic health record (EHR) at the point of prescription. RTPB tools can retrieve cost information, then display therapeutic and/or pharmacy alternatives with lower costs. Health systems can customize how RTPB tools are implemented. As little is known about RTPB tool utilization, we aimed to describe the early experiences, implementation decisions, and outcomes associated with RTPB tools at five U.S. medical centers.

Methods: This is a cross-sectional study at three months after RTPB tool implementation at University of Pittsburgh Medical Center (UPMC), Yale New Haven Health (Yale), Froedtert and Medical College of Wisconsin (F&MCW), Johns Hopkins Health System (JH) and Stony Brook Medical Center (SB). We identified each institution’s implementation characteristics, rates of e-prescriptions where RTPB tools retrieved cost data (retrieval rate), and prescriptions that were adjusted using the tools. Characteristics included whether tool alerts displayed automatically, options for providers to opt out, and savings thresholds to display alternatives. Data sources included outpatient prescription and RTPB data from the EHRs.

Results: Implementation characteristics, RTPB data retrieval rates and prescription adjustment rates at each institution are summarized in table 1.

<table>
<thead>
<tr>
<th>Institution</th>
<th>UPMC</th>
<th>Yale</th>
<th>F&amp;MCW</th>
<th>JH</th>
<th>SB</th>
</tr>
</thead>
<tbody>
<tr>
<td>RTPB Vendor(s)</td>
<td>A</td>
<td>B</td>
<td>A, B</td>
<td>B</td>
<td>B, C</td>
</tr>
<tr>
<td>EHR Vendor</td>
<td>Epic; Cerner</td>
<td>Epic</td>
<td>Epic</td>
<td>Epic</td>
<td>Cerner</td>
</tr>
<tr>
<td>Automatic Display</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Opt-out Option</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Display Threshold</td>
<td>None</td>
<td>None</td>
<td>&gt;$3/mo or $0.10/d</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>RTPB Data Retrieval Rate</td>
<td>49%</td>
<td>8%</td>
<td>60%</td>
<td>38%</td>
<td>20%</td>
</tr>
<tr>
<td>Rx Adjustment Rate</td>
<td>Not Available</td>
<td>1.2%</td>
<td>3.4%</td>
<td>1.4%</td>
<td>0.1%</td>
</tr>
</tbody>
</table>

Discussion: Prior literature has shown that EHR tools can impact cost-related provider ordering behavior2. Rates of RTPB information retrieval across health systems were variable, which may reflect differences in the institutional payer mix, vendor mechanisms for obtaining RTPB information, number of RTPB vendors at an organization and the level of integration of its health system. Prescription adjustment rates were universally low and may be affected by a variety of factors: low RTPB retrieval rates, system design choices in creating meaningful thresholds to prompt adjustments, and implementation challenges such as lack of provider awareness or confidence in these new tools.

Conclusion: RTPB tools represent a paradigm shift in price transparency at the point of medication prescribing. Our early data demonstrate a range of implementation differences across sites. More research is needed to understand factors that affect retrieval of RTPB information, how to optimize tool utilization and their role in patient care.

1 Iuga AO, McGuire MH. Adherence and health care costs. Risk Manag Healthc Policy. 2014;7:35-44
An Interactive COVID-19 Vaccination Appointment Dashboard
Theodore B. Wright, MD, MBI, Charlene Davis, and Connor Rossier, PharmD
Central Virginia VA Health Care System, Richmond, Virginia

Introduction
COVID-19 dashboards have been used for exhibiting rising positivity and mortality rates1, showcase pandemic predictive models2, monitor for outbreaks and trends, and even to inform on returning to work or school3. With the release of the first COVID-19 vaccines, the challenge was to vaccinate as fast possible. The Veterans Affairs’ (VA) prioritized vaccine distribution using a multi-tiered approach4. The Central Virginia VA Healthcare System (CVHCS) provides services to nearly 70,000 unique Veterans yearly, many of which are at high-risk for complications5. Herein, we describe an interactive COVID-19 Vaccine Appointment dashboard meant to track vaccine administration rates as well as scheduling of vaccine appointments.

Methods
Our team maintains an operational reporting platform that allows rapid development and deployment of powerful, interactive visual dashboards for clinical and operational data. Using the VA’s Corporate Data Warehouse (CDW)6, we build tabular data models and interactive user reports with Microsoft’s Power BI software. Development of our dashboard started with the requirements listed in VA’s vaccination campaign and had two primary goals: 1) Identify Veterans in each vaccination cohort, and 2) delineate Veterans by vaccination appointment status. We later expanded on the output. Vaccination cohorts are risk-stratified by age, vaccination, and appointment status by leveraging EHR data. Veteran statuses were categorized for a simplified aggregate graphical display. The final dashboard was used by the CVHCS Incident Command Center (ICC) for review and action.

Results
The dashboard provides local COVID-19 vaccine appointment activity and vaccination rates by cohort (Fig. 1). To facilitate scheduling efforts, vaccination status categories can be selected which instantly display individual Veteran contact information in an accompanying data table. During the first 7 months of 2021, our data shows that, for Veterans with an assigned primary care team at CVHCS, over 50,000 total vaccine doses have been administered and nearly 25,000 of those Veterans are fully vaccinated as of this writing.

Discussion
In addition to usual data reporting, this dashboard’s interactive nature allows specific, actionable data to be located quickly. This key feature allowed us to create a unique tool that helped with facility vaccine scheduling efforts.

Conclusion
Here we demonstrate how visual analytics tools can supplement other information systems such as a scheduling software when new business requirements arise. Our dashboard continues to be used at CVHCS to help vaccinate our Veterans as well as inform clinic staffing and vaccine distribution needs.

References
The Spectrum of Engagement in Chronic Kidney Disease Care at a Large Healthcare System

YiFan Wu, MPH, Adam Wilcox, PhD
Biomedical and Health Informatics, University of Washington, Seattle, WA

Introduction

Chronic Kidney Disease (CKD) has placed a significant burden on society in the USA. Even with advances in information systems like electronic health records (EHR), the recognition rate of CKD remains consistently low. In 2019, researchers reported the recognition rate (both confirmatory lab results and ICD codes in EHR) of CKD was 11.8% among Medicare beneficiaries across ten states in the USA. Besides the low recognition of CKD, insurance coverage and geographical constraints on access to treatment have previously been identified as barriers for patients not seeking specialized care. Similar to HIV care, CKD care can be depicted as a continuum of engagement. In this paper, we assess the CKD recognition rate and the spectrum of engagement in CKD care at the University of Washington healthcare system (UW), as well as identify potential reasons for patients being lost-to-follow-up, where lost-to-follow-up is defined as those who had CKD ICD codes in EHR and clinical manifestation but didn’t seek for specialized care.

Data Sources and Methods

The CKD cohort was identified using EHR in the UW Medicine enterprise data warehouse from 2015 to 2020. We defined the study cohort as following: A. CKD determined by laboratory tests: a patient who had at least two eGFR <60 mL/min/1.73 m<sup>2</sup>, 90 days apart; B. CKD by diagnosis: a patient who received at least one CKD dx of ICD-10 codes N18(1-9); C. Patients who had both lab-confirmed results and dx and at least one renal clinic visit at UW Medicine; D. Maintained CKD condition: patients who met all the criteria above and had at least two UW renal clinic visits(D ⊆ C ⊆ B ⊆ A). We further investigated the patients who had clinical evidence but were missing a CKD dx by examining insurance status. Lastly, we randomly chose 15 patients who had CKD stage >3b (the higher the more severe) and who didn’t have a CKD ICD dx and reviewed the clinical notes in EHR to examine the potential reasons for lost-to-follow-up in CKD care.

Results and Discussion

We identified 39977 CKD patients based on the clinical criteria. 18861 (47.17%) patients had a CKD ICD dx in EHR, and 14900 (79%) of them had insurance coverage. Among the 21116 patients who didn’t have a CKD ICD dx in EHR, only 4434 (21%) of them had insurance coverage on file. Among the patients with CKD dx and abnormal lab values, 7171 (38.02%) had at least one encounter at the nephrology clinic, and 5199 (27.56%) had more than two encounters at the nephrology clinic. Since the insurance status varies significantly between the patients with dx vs. without dx in EHR, we initially thought that not having insurance coverage might be the primary driver of lost-to-follow-up. By conducting clinical notes biopsy, we found that the reasons for lost-to-follow-up can be summarized as 1) patients are getting specialized care from external hospitals out of UW system 2) geographical constraint 3) EHR data quality issue 4) recovered and no longer need renal care. Overall, by focusing on the structured EHR, the recognition rates remain low, that the % of patients without CKD ICD dx raised some concerns. The qualitative study showed that patients do not necessarily fall out of the spectrum of engagement but seek external care.

Conclusion

Our study followed a cohort longitudinally of the spectrum of CKD care in a single large healthcare system in 2015-2020. The study shows the limited engagement of CKD care at a 47.17% recognition rate. Our study shows that integrating both qualitative and quantitative analysis is essential to examine the reasons behind lost-to-follow-up.

References

Development and Validation of a Survival Score for the Emergency Department in Singapore

Feng Xie, BS¹, Bibhas Chakraborty, PhD¹, Nan Liu, PhD¹, Marcus Eng Hock Ong, MBBS, MPH¹,²
¹Duke-NUS Medical School, National University of Singapore, Singapore; ²Singapore General Hospital, Singapore

Introduction

Survival prediction for time-to-event outcome helps treatment determination and resource allocation, especially for patients with critical illnesses or emergencies. At present, survival scores have been pervasively used in healthcare for evaluating patients’ overall survival and represent them as a single indicative score, such as the survival prediction score (SPS) and the Palliative Prognostic Score (PaP). To the best of our knowledge, there is no survival score developed for the Emergency Department (ED), where the patients’ predicted overall survival could provide a valuable reference for clinician’s decision-making. In this study, we aimed to develop a survival score to predict ED patients’ overall survival based on a survival analysis framework.

Methods and Materials

All emergency admission episodes from January 1, 2009 until October 31, 2016 in Singapore General Hospital (SGH) were included in this study. We included demographics, comorbidities and clinical variables. Based on the survival analysis framework, survival time was measured from the time point of patients’ ED admission. Time to death within 365 days after ED admission from all causes was the primary outcome. Our survival score was developed on a novel AutoScore pipeline with machine learning-Random Survival Forest (RSF) for variable selection, Cox proportional hazards regression for score weighting and Kaplan–Meier methods for outcome analysis. The recent ED episodes in 2016 were assigned to the testing cohort for performance evaluation and comparisons. We compared it with the other scores such as Modified Early Warning Score (MEWS), National Early Warning Score (NEWS), Score for Emergency Risk Prediction (SERP) by applying concordance (C-index, D-index) and time-dependent receiver operating characteristic (ROC) analysis.

Results and Evaluation

A total of 337,000 unique emergency admission episodes were included. Seven components were selected by RSF to form the survival score (Table 1), consolidating patients' predicted overall survival into one single indicative score. The derived triage score outperformed several baseline scores (Table 2). It significantly predicted and stratified patients’ survival on the Kaplan–Meier Curve (Figure 1). Age and comorbidities were powerful predictors for patient’s overall survival in the aging population. Our survival score outperformed other comparators on various metrics.

Conclusion

We developed the survival score from EHR data of eight-year ED admission episodes by combining machine learning and the regression analysis. This score could help anticipate the patient’s overall survival when patients are admitted from the ED.
Risk Stratifying Heart Failure Readmissions of VAD-Eligible Patients Using Laboratory Data

Jinchen Xie, MS¹, Arman Kilic, MD², Rema Padman, PhD¹
¹Carnegie Mellon University, Pittsburgh, PA, United States; ²Medical University of South Carolina, Charleston, SC, United States

Heart failure represents a spectrum of disease severity with specific clinical criteria for advanced therapy, including heart transplantation and left ventricular assist device (LVAD) implantation. However, the transition between phases of HF severity and defining the timing for advanced therapies are subjective, lacks standardization and a data-driven approach, resulting in a missed opportunity for potentially impactful interventions. We hypothesize that a better understanding of the latent heterogeneous groups in the VAD-eligible cohort will lead to more accurate and clinically interpretable risk models for intervention timing and choice of therapy. In this study, we investigate this heterogeneity in HF patients with multiple readmissions by risk stratifying the population using routinely collected laboratory data during inpatient stay.

We applied multi-trajectory modeling² using six laboratory tests at each visit, across up to ten consecutive admissions, to identify distinct subtypes of HF patients eligible for VAD therapy and examine the differences across the resulting groups. We excluded patients from the study cohort who did not meet the criteria to receive VAD therapy at their initial encounters. Using the six laboratory tests, including estimated glomerular filtration rate (eGFR), sodium, hemoglobin, potassium, platelets, and white blood cell counts, we estimated a parsimonious, group-based, multi-trajectory model and evaluated multiple fitness criteria to cluster patients into an optimal number of distinct risk groups. Eleven risk factors established at the index admission, such as age and gender, were also included in the model as baseline covariates. Standard significance tests evaluated differences in key characteristics between groups.

The final analysis cohort included 1547 patients identified as VAD-eligible at their index admission. We identified five distinct trajectory groups evolving over multiple admissions using the six types of laboratory data (Figure 1). Groups 1 includes the most severely ill HF patients with the highest all-cause 30-day and 1-year mortality rates and the highest proportion with a history of multiple comorbidities. These patients also started at the lowest eGFR level compared to the other four groups. Group 2, the second-highest risk group, had the most patients on Anti-Platelet (P2Y12) and Class I Anti-arrhythmic prescription. Group 4, the lowest mortality risk group, had relatively young patients and the lowest proportion of multiple comorbidities and clinical events history. Group 3 and Group 5 formed an interesting comparison. While the two groups had similar average age and comorbidities, Group 3 patients, on average, received VAD just within the first year after index admission and had a lower all-cause mortality rate by the end of follow-up compared to those in Group 5 who received VAD, on average, several months later.

The group-based multi-trajectory model shows promise in risk stratifying patients with multiple HF readmissions into distinct HF severity groups using underutilized, routinely collected laboratory data and baseline covariates. The model also revealed eGFR’s impact as a dominant factor in identifying the higher mortality risk groups of VAD-eligible heart failure patients. We need to further examine the differences in the timing of VAD implantation across groups, and the adverse events occurring along patients’ trajectories to identify patterns in each group, both of which could help guide the selection of patients for early interventions.

References
Hand hygiene monitoring by positioning technology utilizing IoT devices

Keiko Yamashita¹, Shintaro Oyama¹, Satoshi Yamashita¹, Chiaki Funada¹, Kikue Sato¹, Taiki Furukawa¹, Aki Sugano¹, Daisuke Kobayashi¹, Hiroshi Tomozawa³, Yuji Sakamoto⁴, Yoshinori Ideno⁴, Kensaku Mori⁵, Yoshimune Shiratori¹
¹Medical IT Center, Nagoya University Hospital, Nagoya, Japan; ²Division of Medical and Healthcare Systems, Kobe University, Kobe, Japan; ³SATO HEALTHCARE CO., LTD., Tokyo, Japan; ⁴CARECOM Co., LTD., Tokyo, Japan; ⁵Graduate School of Informatics, Nagoya University, Nagoya, Japan

Introduction
Hand hygiene (HH) is important for infection prevention in the medical field, and WHO has been successful in preventing nosocomial infections by organizing the timing when HH is necessary in the patient care. Typical monitoring methods are the direct observation method and the hand sanitizer usage measurement method, but the problem is that it takes time and labor. In this study, we investigated whether HH monitoring using the Internet of Things (IoT) could establish a screening method for direct observation and a verification method for indications at each facility.

Methods
In the standard ward of Nagoya University Hospital, we have introduced a high-precision positioning system with an accuracy of 10 to 30 cm using the AOA (Angle of Arrival) method of BLE4.0 (Quuppa Intelligent Locating System, Quuppa Oy, Finland). The BLE locators (26 units) were placed on the ceiling at a height of 3 m in the center of the hospital room. We attached a BLE beacon to the pump part of my personal hand sanitizer. When hand sanitizer is pushed, the BLE signal transmitted from the pressed beacon is detected by the nearest locator (Figure 1). Two nurses were conducted in this study and surveyed for one week.

Results
The percentage of HH practices that defined HH practices by timing required for direct observation was 55.2% within 30 seconds of admission, 42.8% within the bed area, and 2.0% within 30 seconds after leaving the room. The HH execution rate was high, as the staying hour was longer (Figure 2).

Discussion
There is also a method of using images as an automatic HH monitoring system, but in order to replace the direct observation method, the timing of HH implementation is important, and we thought that location information was necessary for that purpose. Also, although it is difficult to identify the subject using the direct observation method, the IoT HH monitoring method can regularly observe the number of visits to the room by pump sensor and the HH implementation status at each timing, and is considered to be useful for improving effective HH implementation. In addition, it is effective in suppressing horizontal transmission of infection and may contribute to infection control, and it is considered to be one of the screening methods of direct observation, which is a conventional monitoring method.

Conclusion
From this research, HH monitoring using IoT will be an excellent future innovation through important technology for controlling nosocomial infection control.

References
Understanding Algorithmic Bias in Clinical Prediction Models

Mengying Yan, MS\textsuperscript{1}, Michael J. Pencina, PhD\textsuperscript{1}, Benjamin A. Goldstein, PhD\textsuperscript{1}

\textsuperscript{1}Department of Biostatistics & Bioinformatics, Duke University, Durham, NC, USA

Introduction

Increasing attention has been drawn to the problem of racial disparity in clinical prediction models (CPMs). Some have suggested that one shouldn’t use race in CPMs\textsuperscript{1}. We characterize the problem and discuss how accounting for a grouping variable such as race in the model would affect algorithmic bias.

Methods and Results

We define algorithmic bias in a CPM via differential miscalibration, i.e. the model differentially under (or over) predicts risk for a particular group. Moreover, we frame the creation of such bias via a differential missing data process, where conditional on having the outcome of interest (e.g. asthma exacerbation) the outcome is differentially observed (i.e. via coming to the hospital). In such a scenario, we posit that one may learn a CPM that perpetuates algorithmic bias.

We sought to investigate how the inclusion of a grouping variable into a CPM could affect such bias. We performed a series of simulations, simulating a two stage process where the grouping variable may impact either the risk of the true outcome (stage 1) or probability of observing the outcome (stage 2). Suppose $X_1$ is any risk factor, $X_g$ is the grouping variable, $E$ is the true outcome and $Y$ is the observed outcome. The two stages are:

\[
\text{logit}(P(E = 1)) = \beta_0 + \beta_1 X_1 + \beta_g X_g
\]

and

\[
\text{logit}(P(Y|E = 1)) = \gamma_0 + \gamma_g X_g
\]

Here, $X_g$ potentially affects the true risk through $\beta_g$, and potentially affects probability of observing the outcome through $\gamma_g$. After simulating the true data, we learned a CPM, where both included and did not include $X_g$ as a predictor. We then compared learned risk to the true risk.

The results are shown as calibration plots. In (a) and (b), grouping variable does not impact stage 1 but does impact stage 2. Here including $X_g$ in the model does induce more algorithmic bias (b) Conversely, in (c) and (d), $X_g$ impacts the true risk of the outcome (stage 1) but not the observability (stage 2). Here ignoring it in the CPM (c) leads to algorithmic bias where including it leads to no algorithmic bias (d).

![Calibration plots](image)

Figure 1: True versus predicted risk by grouping variable $X_g$. $X_g$ impacts stage1 or stage2, including or excluding $X_g$ in model

Conclusion

We show that if the grouping variable impacts the true outcome, then including it reduces algorithmic bias. However, if the grouping variable only impacts the observation process, it is better to exclude it in model. Therefore, the advice to simply drop race from clinical prediction models\textsuperscript{1} may be too simplistic.

References

Implementation of Medication Alerts to Reduce Wrong-Drug and Wrong-Patient Errors in CPOE Systems

Yuyang Yang, BS1, David Liebovitz, MD2, Bruce Lambert, PhD3, Bill Galanter, MD, PhD4, Jason Adelman, MD5, Thomas Byrd IV, MD6

1-3,6Northwestern University - Feinberg School of Medicine, Chicago, IL; 4University of Illinois at Chicago, Chicago, IL; 5Columbia University Medical Center, New York, NY

Introduction
Wrong-drug (WD) and wrong-patient (WP) errors occur at a rate of roughly one per thousand orders in inpatient and outpatient settings, resulting in millions of errors annually that can cause severe harm to patients1. Prescriber self-interception of errors can be measured using EHR system logs to detect two types of events: 1) abandon-and-reorder (AAR), where the prescriber starts and abandons an incorrect order before signing it, and then reorders for the correct drug or patient and 2) retract-and-reorder (RAR), where the prescriber cancels an incorrect order within minutes of signing, and re-orders for the correct drug or patient. Prior work by our collaborators found that creating indication alerts to notify prescribers when an ordered medication does not match the patient’s current problems (e.g., metformin order in absence of diabetes) was able to prevent errors2,3. This poster describes the implementation of indication alerts for the Northwestern branch of a multisite follow-up study to determine the effect of indication alerts on prevention of wrong drug/patient errors and addition of problems to patient problem lists during clinical encounters.

Methods
Indication Alerts - Alerts are delivered to clinicians during medication ordering via pop-up dialog (Figure 1). Alerts prompt prescribers to place diagnoses on the problem list when medication orders are initiated if no justification exists.

Identification of WD/WP errors - We used the electronic data warehouse at Northwestern to extract drug name, prescriber’s name, time of medication selection, patient identifiers, and content of problem lists at time of medication selection. We developed R scripts to identify WD/WP errors based on the Wrong Drug Retract and Reorder (WD-RAR) tool, a National Quality Forum (NQF Measure #2723) validated measure.

Results and Conclusion
As of submission of this manuscript, we have finished design of all indication alerts and have collected pre-implementation data on rates of WD/WP errors in our hospital system. We have recently turned-on indication alerts and are actively collecting post-implementation impact on error rates and problem list completion.

References
Using multivariate models to examine the impact of COVID-19 pandemic and gender differences on health and health care

Jiancheng Ye1*, Zhimei Ren2
1 Feinberg School of Medicine, Northwestern University, Chicago, IL, USA; 2 Department of Statistics, University of Chicago, Chicago, IL, USA

Introduction
The coronavirus disease 2019 (COVID-19) pandemic has uprooted conventional health care delivery for various care, such as routine ambulatory care and mental health care, requiring health systems to rapidly adopt new capabilities. With mandated social distancing policies in place during the COVID-19 pandemic, health care providers have been forced to get closely acquainted with virtual health. We drew data from the National Cancer Institute's 2020 Health Information National Trends Survey (HINTS). We described and compared the characteristics of social determinants of health, physical activity, mental health, alcohol use, and patterns of social networking service use and health information data sharing. All analyses were weighted to provide nationally representative estimates.

Methods
To understand the joint effect of gender and pandemic, we characterize the response of interest via multivariate models: for continuous responses, linear models are considered and for discrete responses, generalized linear models are considered. In the models, we include the effect of gender, the pandemic, and the interacting effect of gender and pandemic, along with demographic covariates to account for the potential confounding factors. Data used in this study were from the fourth round of data collection for HINTS 5 (Cycle 4), which began in February 2020 and concluded in June 2020. Because data collection for Cycle 4 started before COVID-19 became an international pandemic and continued after the pandemic was declared by the World Health Organization, the variable PANDEMIC was created to flag respondents whose survey was received after the COVID-19 pandemic declaration (March 11, 2020). This variable will facilitate the examination of responses before and after the COVID-19 pandemic that became a widespread issue of concern in the United States. The final HINTS 5, Cycle 4 (2020) sample consists of 3,865 respondents. Note that 73 of these respondents were considered partial completers who did not answer the entire survey.

Results
For the characteristic made appointments with a healthcare provider online, the effect of sex before the pandemic is not significant but becomes significant after the pandemic, where the aOR between the female and male population is e^{-0.36}=0.70 (this means the odds of making appointments with a health care provider online for females is 0.7 times that of male); the effect of the pandemic is not significant either before or after the pandemic, and the interacting effect of pandemic and sex is not significant either. As for looking for health or medical information for oneself, the effect of sex is significant before the pandemic, with an aOR between the female and male population being e^{-0.65}=0.52; the effect of sex is also significant after the pandemic, with an aOR between the female and male population being. Similarly, we see a significant effect of sex both before and after the pandemic in characteristics including look up medical test results, share health information on social networking sites, participate in an online forum or support group, visit a social networking site, use smartphone to track progress on a health-related goal and use an electronic wearable device to monitor or track your health or activity; we also see a significant effect of sex after the pandemic with regard to the characteristics have “apps” related to health and wellness.

Discussion
The COVID-19 pandemic has devastated communities of marginalized populations, exposing the deep inequities of the health care system. This study used multivariate models to understand the joint effect of gender and pandemic on mental health, physical health, and behavioral health. The findings of this study demonstrate that significant inequities are also present among patients in accessing necessary health care and relevant information. Responses to the pandemic should consider diverse perspectives, including sex and gender. Decisions that are informed by accurate data and include a gender perspective are more likely to be effective.

References
Predicting-30 Day All-Cause Unplanned Hospital Readmission of Patients That Developed Acute Kidney Injury During Intensive Care Unit Stay

Zaid K Yousif, PharmD, MAS1, Michael Hogarth, MD1, Linda Awdishu, PharmD, MAS2

1Division of Biomedical Informatics, Department of Medicine, University of California, La Jolla, San Diego (UCSD), California, USA; 2Division of Clinical Pharmacy, UCSD, Skaggs School of Pharmacy and Pharmaceutical, La Jolla, California, USA

Introduction: Hospital-acquired acute kidney injury (HA-AKI) is a frequent adverse drug event with short- and long-term complications. HA-AKI affects up to 50% of patients during their first week of intensive care unit (ICU) stay. It is estimated that 1 in 3 HA-AKI survivors will be readmitted to the emergency department or die within 30 days of hospital discharge. In this study, we describe developing a 30-day unplanned hospital readmission risk prediction model of patients who developed HA-AKI during ICU stay using electronic health record (EHR) data.

Methods: We conducted a retrospective cohort study using the Multiparameter Intelligent Monitoring in Intensive Care (MIMIC) IV database. Patients who developed HA-AKI during ICU stay meeting the 2012 Kidney Disease: Improving Global Outcomes Clinical Practice Guideline (KDIGO) criteria were included. The Centers for Medicare & Medicaid Services (CMS) criteria for readmission measures was adopted. Readmission was defined as all-cause hospitalization within 30 days from the index hospitalization. Per CMS guidelines, planned hospitalizations were not counted as readmissions. A total of 1136 potential variables were evaluated. To avoid overfitting, we excluded 712 variables with < 2% prevalence or > 60% missing, leaving 424 variables. Feature selection was performed using L1 (least absolute shrinkage and selection operator—LASSO) logistic regression. This resulted in the selection of 27 variables which included: demographics (1), healthcare utilization and socioeconomic factors (9), inpatient medications (7), labs (4), diagnoses (4), and procedures (2). An unpenalized logistic regression model was constructed with 30-day unplanned readmission as the outcome and the 27 variables above as predictors. The area under the receiver operator characteristic curve (AUC) was used to assess the model’s discrimination. Additionally, we compared our model discrimination to LACE index, a validated risk prediction score for hospital readmission.

Results: After applying the eligibility criteria, a total of 22,214 index hospitalizations were included. There were 4,217 (19%) 30-day unplanned hospital readmissions. Our model moderately predicted all-cause unplanned readmissions (AUC 0.66, 95% CI 0.64-0.68), and significantly outperformed LACE score (AUC 0.57, 95% CI 0.54-0.59). In general, variables reflecting increased healthcare utilization and unfavorable socioeconomic status increased the risk of readmission. Clinical findings associated with a decline in kidney function increased the risk of readmission.

Discussion: As the rates of HA-AKI survivorship increase, interventions to reduce unplanned readmission targeting this population are needed. To the best of the authors’ knowledge, this is the first effort to develop a 30-day unplanned hospital readmission risk prediction model of patients who developed HA-AKI during ICU stay. When designing our model, we included modifiable risk factors whenever possible. These variables can be utilized to guide the development of clinical interventions to identify and target high-risk patients following hospital discharge. Future work includes validating the model using an external dataset.

References:
Electronic Health Records Reveal Statins Prescription Trends among Small to Medium Sized Practices

Jingzhi Yu1, Ann Andee Wang1, Huyen Vu1, Nicholas Soulakis1, Yacob Tedla1, Abel Kho1
1Northwestern University Feinberg School of Medicine, Chicago, IL

Introduction: Statins have been found to be very effective in reducing cardiovascular morbidity and mortality. While the 2013 ACC/AHA cholesterol management guidelines expanded criteria for patients eligible for moderate to high intensity statins therapy, prior research has found adoption rates leave much room for improvement. In our study, we examine the breakdown of prescribed statins dosage strength and the change across time in small to medium sized practices. These practices account for over 50% of primary care practices in the US; however, prior research in this patient population is lacking.

Methods: Our data was extracted from the EHRs of practices that participated in an Agency of Healthcare Research and Quality (AHRQ)-funded initiative, Healthy Hearts in the Heartland (H3), which aimed to improve quality through practice facilitation. The initiative obtained patient level data from 2013 to end of 2017. Statins prescription was ascertained by the patient’s current medication list or medication prescription table. Dosage was parsed from the descriptions of the medications prescribed. Based on actual numeric dosage, statins were categorized into low, moderate, and high intensity, according to the 2013 guideline. At each month, the number of unique patients was accumulated from the previous months and the percentage of each class of statins was calculated for each month. At any given month, patients with more than one class of statins prescribed leading up to that month were attributed with the highest class of statins prescribed for them. Furthermore, we compared the trends in independent practices and federally qualified health centers (FQHCs), which provide care to a different patient population.

Results: Our dataset included 48,059 unique patients from 70 practices in the northern Midwest U.S. Over the course of 4 years, we saw the percentage high intensity statins increase in the total patient population that were prescribed statins. FQHCs accounted for 33 practices and 20,849 patients. We saw a corresponding reverse trend in the percentage of low intensity statins, while the percentage of moderate intensity statins remained the same. When comparing independent practices and FQHCs, we found greater changes in prescription behavior in FQHCs. Independent practices saw a smaller increase in high-intensity statins, and a constant percentage for low-intensity statins.

Discussion and Conclusion: We found growth in high-intensity statins prescription in small to medium sized practices. However, FQHCs were found to be more likely to enact change in prescription habits compared to independent practices. Further research should be conducted to understand this discrepancy to guide future quality improvement work in small- and medium-sized practices.

Fig 1. Statins intensity trends for all practices (left), FQHC practices (middle), independent practices (right).

Prediction of COVID-19 Case Severity Using Synthetic Data Derived from the National COVID Cohort Collaborative

Noa Zamstein, Ph.D.¹; Andrew J. Neumann, M.B.A.²; Randi Foraker, Ph.D., M.A.³; Jason A. Thomas, B.S.⁴; Adam Wilcox, Ph.D.⁵; Jon D. Morrow, M.D., M.A., M.B.A.⁶,⁷

¹MDClone Ltd., Be’er Sheva, Israel; ²Univ. of Colorado, Aurora, CO; ³Washington Univ., St. Louis, MO; ⁴Univ. of Washington, Seattle, WA; ⁵New York Univ., New York, NY

Introduction: Broad access to data facilitates computational reporting capabilities during a pandemic. Such data are used to quantify daily tests, infections, hospitalizations, and deaths. The National COVID Cohort Collaborative (N3C) of the National Institutes of Health (NIH) brings together longitudinal patient-level data across multiple institutions from different geographic regions into a centralized repository. Synthetic data have been validated as functionally equivalent to original data for analyses.²,³

Objective: To demonstrate prediction of hospital severity and course of COVID-19 illness at admission using a synthetic data set, complementary to a published demonstration of N3C data capabilities.⁴

Methods: We included all adults from the N3C database who tested positive for SARS-CoV-2 who were admitted to hospital between March 2020 and July 2021. Patients were included if they had a hospitalization lasting for at least 24 hours and had at least one additional laboratory result recorded within 24 hours of admission. We defined severe cases as those involving mechanical ventilation or extracorporeal membrane oxygenation (ECMO) or ending in death or discharge to hospice. We included laboratory measures, vital signs, background conditions and demographics in the models. Data were extracted and synthesized using the MDClone ADAMS Platform [MDClone Ltd., Be’er Sheva, Israel]. The MDClone Synthetic Data Engine was used to create a synthetic data set that maintains the statistical properties and inter-relationships of the original data. We developed machine-learning models for predicting the maximum severity during the admission, including over 40 variables. We trained the models against both the original data and the synthetic derivative, and we tested the models against the same data set.

Results: 181,159 patients met the inclusion criteria, of whom 16.1% had severe disease. Median length of stay was 5 days for non-severe and 12 days for severe cases. Non-severe cases had a mean of 2.0 comorbidities and severe cases had 3.4. The area under the curve (AUC) and precision-recall AUC (PR-AUC) of the gradient-boosted-machines (GBM) models demonstrate that training on the synthetic data set achieves the same high degree of utility and predictiveness as the model trained on the original data (Fig. 1a). The feature importance of the predictors in the models trained on the original and synthetic data were highly concordant (Fig. 1b).

Discussion and Conclusions: Synthetic data can represent patient characteristics and support advanced predictive analytics with statistically equivalent results as an original data set. We used these synthetic data to train a non-linear model incorporating a very large number of variables, and we achieved similar results to models developed with the original data. Using synthetic data, we successfully studied hospitalized patient characteristics on the first day of admission and predicted disease severity. Our study was limited by the broader challenges of using data in a multi-institutional cohort. Use of synthetic data can make data accessible to “citizen scientists,” enabling a broader audience to generate pandemic-related knowledge. Future centralized data networks may benefit from the use of synthetic data to avoid the sharing of real human-subject data while maximizing utility.

References

Fig. 1a (left): AUC and PR-AUC for GBM performance on original and synthetic data. Fig. 1b (right): Feature importance for the GBM model.
Improving COVID-19 Vaccine Uptake of Mental Illness Patients using Learning Health System Framework

Xiaoming Zeng, MD, PhD, Austin Hall, MD, Victoria Jeffries, MSPH, Stacey Burgin, MA, John Gilmore, MD, Beth Rossi, BA, Thava Mahadevan, MS
Department of Psychiatry, University of North Carolina at Chapel Hill

Introduction
The nationwide rollout of COVID-19 vaccines has accelerated the pace of mitigating the pandemic. It is critical to have most of the population inoculated with the vaccines for achieving the "Herd Immunity." Studies using national data found that patients with existing mental illnesses are more susceptible to COVID-19 infections.\(^1\)\(^,\)\(^2\) As a result, people with severe mental health are advocated to receive the COVID-19 vaccine as a high-priority group.\(^3\) However, patients with mental illness often lack the needed resources and means to complete the vaccination process. Given the underlying mental disorders, they may also exhibit hesitancy for vaccinations. Therefore, this group requires extra help to overcome the barriers to getting vaccinated.

Description
Using the Learning Health System (LHS) framework, we report an ongoing quality improvement project to improve the uptake of COVID-19 vaccines for patients in the Center for Excellence in Community Mental Health at the University of North Carolina. LHS is defined by the Institute of Medicine as a system in which "science, informatics, incentives, and culture are aligned for continuous improvement and innovation, with best practices seamlessly embedded in the delivery process and new knowledge captured as an integral by-product of the delivery experience." A common framework to present LHS is repeated closed circles consisted of multiple steps. Figure 1 represents one of such cycles, and the steps of the project are specified below:

1. Assemble of data: We synthesize data from multiple sources to support the project – EHR, Master Patient Registry, Patient Feedback Tracker Form, and others.
2. Analyze the data: We use a data visualization tool to show the weekly progress of vaccinations among all eligible patients. We reported KPIs of proportions of patients receiving zero, first, or second dose.
3. Interpret the results: The visualization and progress of patient engagement are discussed at weekly steering committee meetings. New knowledge and insights often are gained during this step.
4. Feedback: Tailored messages are created for different groups – attendings, residents, therapists, office managers – to guide their activities in the next step.
5. Change: Changes are implemented at the end of each cycle based on steps 1-4. For example, the project team quickly realized that an increased number of patients have their second doses overdue. The workflow was adjusted accordingly to contact these patients to re-engage them into the process immediately.

Lessons learned
Multiple lessons have been learned from the ongoing project:

1. Not all data are readily available for tracking the vaccination process. Our EHR system does not capture all vaccine data for all patients. If a patient receives a vaccine dose outside our health system and does not have recent appointments, then that data is not captured. As of today (3/9/2021), the state vaccine tracking system has not been able to share data with the local EHR system yet. Some of the data, such as patients' readiness and hesitancy, must be collected and processed by contacting patients directly, becoming time-consuming, error-prone, and costly.

2. Engaging stakeholders is critical. The project involves multiple stakeholders in our center – physicians, therapists, clinic managers, social workers, and even patients. The project’s leadership frequently communicates with the different stakeholders to inform them of the latest status of the vaccination among patients. Buy-in from and quick action of all involved is critical for the success of the project.

References
Developing an Ontology for Social and Behavioral Determinants of Health
Hansi Zhang, MS\textsuperscript{1}, Xi Yang, PhD\textsuperscript{1}, Thomas J George, MD, FCAP\textsuperscript{2}, William Hogan, MD, MS\textsuperscript{1}, Jiang Bian, PhD\textsuperscript{1}, Yonghui Wu, PhD\textsuperscript{1}

\textsuperscript{1}Departments of Health Outcomes and Biomedical Informatics, College of Medicine, University of Florida, Gainesville, Florida, USA
\textsuperscript{2}The Division of Hematology and Oncology, College of Medicine, University of Florida, Gainesville, Florida, USA

**Introduction:** There is an increasing interest in examining social and behavioral determinants of health (SBDoH) in health outcomes research to understand their roles in shaping people’s health and reduce potential confounding issues and misclassification errors. However, SBDoH information is captured in heterogeneous data sources and is rarely being effectively used. One potential source for SBDoH is the clinical narratives. To facilitate the extraction of SBDoH information from clinical narratives, we developed a formal ontology for SBDoH with 7 high-level classes and 38 subclasses to help organize SBDoH in electronic health record (EHRs). This ontology will be used for developing information extraction pipelines to help organize and normalize SBDoH information from clinical notes in a consistent way with standardized and shared vocabularies.

**Methods:** We started with the Basic Formal Ontology (BFO), a well-established upper-level ontology, as the basis. Following best practices in ontology engineering, we aim to define the entities (classes and relations related to SBDoH), reusing and adapting existing ontological resources as much as possible. We systematically reviewed concepts and terms relevant to SBDoH through literature search such as the World Health Organization (WHO), Healthy People 2020, and Centers for Disease Control and Prevention (CDC), as well as existing knowledge of SBDoH such as ICD-10 and NCI thesaurus. Based on these concepts and terms, we identified key classes and relations with domain experts. On the other side, we recruited domain experts to annotate SBDoH concepts captured in clinical notes from the University of Florida (UF) Health system. The classes developed in the SBDoH ontology were used to guide manual annotation. The annotation results were used to verify and improve the ontology in an iterative developing procedure. Currently, we are working on identifying and reusing classes and relations in existing ontologies to represent SDoH (i.e., a review of existing widely accepted ontologies will be conducted using Ontobee \cite{7} to find relevant classes and relations) while maintaining clear semantics. When there is none that can be reused or there are new SBDoH concepts identified from UF health clinical notes, we will create new classes and relations in the hierarchy. Following our annotation procedure, this ontology will be iteratively developed to support organizing the SBDoH in real-world electronic health records.

**Results and Conclusion:** We developed an SBDoH ontology with 7 main classes and 38 subclasses. Figure 1 shows the architecture for organization SBDoH related classes in SBDoH ontology. We verified and further improved this ontology through annotating SBDoH concepts from a total number of 500 clinical notes. The proposed ontology will be a valuable resource to help organize SBDoH concepts in EHRs.

**Figure 1.** Classes and subclasses identified in SBDoH ontology.

![Ontology Diagram](image.png)

**Acknowledgement:** This study was supported by a Patient-Centered Outcomes Research Institute® (PCORI®) Award (ME-2018C3-14754), and a grant from the National Cancer Institute, 1R01CA246418 R01.

**References:**

Predicting Next-Day Discharge via Electronic Health Record Audit Logs

Xinneng Zhang*, BS1, Chao Yan*, MS1, Mayur B. Patel, PhD2,3, Bradley A. Malin, PhD1,3, You Chen, PhD1,3

1Vanderbilt University, Nashville, TN; 2Veteran Affairs Tennessee Valley Healthcare System, Nashville, TN; 3Vanderbilt University Medical Center, Nashville, TN (* First authors)

Abstract

EHR audit logs data capture clinicians’ interactions with patient records and provide clues into clinicians’ assessments and insights into patients’ clinical status. It has been utilized to support many endeavors, including clinical task analysis, time-motion studies of clinician activities, and care structure analysis. Machine learning models trained on the audit log data can significantly improve hospital-wide next-day discharge prediction performances, which provides an excellent opportunity for more efficient hospital capacity management.

Introduction

Hospital capacity management depends on accurate real-time estimates of hospital-wide discharges. [1] This study aims to support next-day discharges predictions with EHR audit log data, which is a resource that captures EHR users’ granular interactions with patients’ records. [2]

Method

This study focused on the EHR data for all adults admitted to Vanderbilt University Medical Center in 2019. We learned three light gradient boosting machine (LGBM) models [3] to 1) assess the value that EHR audit log data adds to the daily prediction of discharge likelihood within 24 hours and 2) compare different representation strategies. We applied Shapley additive explanations (SHAP) [4] to identify the most influential types of user-EHR interactions for discharge prediction.

Results

The data includes 26,283 inpatient stays, 133,398 patient-day observations, and 819 types of user-EHR interactions. The model (N-INT) using the count of each type of interactions in the recent 24 hours and other commonly used features, including demographics and admission diagnoses, achieved the highest area under the receiver operating characteristics (AUROC) curve of 0.921 (95% CI: 0.919-0.923). The model (ONLY-INT) that only includes user-EHR interactions also achieved a high AUROC of 0.890 (0.888-0.892). By contrast, the model (NO-INT) lacking user-EHR interactions achieved a significantly worse AUROC of 0.862 (0.860-0.865) than those with interactions (p-value <0.001). Among the 20 most influential factors (in terms of SHAP values) of the best performing model, 50% were user-EHR interaction features with distinct semantics.

Conclusion

EHR audit log data contains rich information such that it can improve hospital-wide discharge predictions. This study is limited that it focused on the raw sequence of user-EHR interactions and did not relate them to the corresponding clinical tasks.

References

CCF-CL: Forecasting the Clinical Status of a Patient Through Contrastive Learning

Ziqi Zhang, BS¹, Chao Yan, MS¹, Xinmeng Zhang, BS¹, Steve L. Nyemba, MS²,
Bradley A. Malin, PhD¹,²
¹Vanderbilt University, Nashville, TN; ²Vanderbilt University Medical Center, Nashville, TN

Objective

An individual’s health status changes over time, such that it is useful to forecast what, if any, health problems an individual is at risk of developing. Over the past several years, deep learning models have made significant progress in clinical concept forecasting. However, the efficacy of these models is limited in practice, particularly when they are trained with a small case group, which is a typical problem due to the low prevalence of many diseases. In this paper, we introduce CCF-CL, a framework to support clinical concept forecasting in the face of a limited number of positive training instances.

Method and materials

CCF-CL is composed of two components. The first component, which possesses almost all parameters in the model, is trained to learn informative patient representations in an unsupervised manner, with ground-truth labels not involved. The second component, which is composed of a limited number of parameters, is trained to fit the derived patient representation with ground-truth labels. In doing so, the training process is encouraged to neglect the labels to the greatest extent possible and, thus, focus on the relationships between features. Figure 1 illustrates the traditional supervised paradigm and CCF-CL.

To enable such a learning paradigm, we introduce contrastive learning into the representation modeling of EHR data. We also develop a data augmentation strategy specifically to adapt contrastive learning into EHR modeling. We evaluate the CCF-CL’s performance by comparing it to current state-of-the-art model architectures trained in the supervised manner, with EHR data from 48,000 and 16,000 patients, each with more than 25 distinct healthcare episodes, from Vanderbilt University Medical Center and the All of Us research program, respectively. Specifically, we perform forecasting for 181 and 120 clinical concepts (each corresponds to a CCS diagnosis code) selected from the two data sources.

Result

Our experiments indicate that the average performance improvement, in terms of area under the receiver operating characteristic (AUROC) curve, for clinical concept forecasting tasks is 0.045 (7.8%) and 0.067 (10.9%) for VUMC and AOU data, respectively. We also found that the improvement is particularly pronounced when for clinical concepts where the proportion of positive to negative cases is large.
Deep Learning Approaches for Breast Cancer Characteristics Extraction from Electronic Health Records

Sicheng Zhou, MS¹, Liwei Wang, MD, PhD⁵, Nan Wang⁵, Sunyang Fu, MHI⁵, Chetan Shenoy, MD⁴, Anne Blaes, MD⁴, Hongfang Liu, PhD⁵, Rui Zhang, PhD¹,²

¹ Institute for Health Informatics, ²Department of Pharmaceutical Care & Health Systems, ³School of Statistics, ⁴Department of Medicine, University of Minnesota, Minneapolis, MN, USA; ⁵Department of Artificial Intelligence and Informatics, Mayo Clinic College of Medicine, Rochester, MN, USA

Introduction

Breast cancer is one of the most prevalent and lethal cancers for women in the US. Precision medicine and related translational research need the support of a large amount of cancer specific clinical information of patients. Effective and automatic extraction of the target information from clinical narratives of EHR remains an important research topic. The objective of this study is to develop and evaluate a set of NER algorithms, i.e., conditional random fields (CRF), bidirectional long short-term memory + CRF (BiLSTM-CRF) and Bidirectional Encoder Representations from Transformers (BERT) fine tuning models, to extract breast cancer characteristics from clinical narratives in EHR.

Methods

The data used in this study was obtained from the EHR of the University of Minnesota Clinical Data Repository. 200 pathology reports and clinical notes of breast cancer patients were randomly sampled and annotated by two annotators. The annotation followed the BIO format. The extraction of breast cancer characteristics from texts can be framed as an NER task. The CRF model were developed as the baseline model. The specific features used for CRF include the length of each token and the pre-trained word embeddings of each token. BiLSTM-CRF models were developed to finish the NER task. The input features for the BiLSTM-CRF model were 6 different pre-trained word embeddings. 3 of them are Word2Vec word embeddings while 3 of them are GloVe word embeddings. And in this study, we fine-tuned and evaluated the BlueBERT pre-trained model on our NER task. The BlueBERT model was pre-trained on PubMed abstract and public available clinical notes, i.e., MIMIC-III that contain over 4500 million words, which is suitable for transfer learning task in clinical domain [1].The BlueBERT model was pre-trained on PubMed abstract and public available clinical notes, i.e., MIMIC-III that contain over 4500 million words, which is suitable for transfer learning task in clinical domain. We applied the name entity level evaluation for all NER models. 70% of the annotated data was used as training set, 10% was used as development set and 20% was used as test set. F1 score was used as evaluation metric. We evaluated the performance using both exact match and lenient match on various name entities.

Results

Table 1—NER evaluation measured by exact and lenient match F1 scores (in parenthesis) for all models.

<table>
<thead>
<tr>
<th>Entity Number</th>
<th>CRF</th>
<th>BiLSTM-CRF</th>
<th>CRF</th>
<th>BiLSTM-CRF</th>
<th>CRF</th>
<th>BiLSTM-CRF</th>
<th>CRF</th>
<th>BiLSTM-CRF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Receptor type</td>
<td>1607 (29)</td>
<td>0.977 (0.978)</td>
<td>0.984 (0.989)</td>
<td>0.990 (0.993)</td>
<td>Cancer grade value</td>
<td>256 (15)</td>
<td>0.692 (0.693)</td>
<td>0.918 (0.920)</td>
</tr>
<tr>
<td>Receptor status</td>
<td>386 (14)</td>
<td>0.503 (0.504)</td>
<td>0.869 (0.870)</td>
<td>0.956 (0.957)</td>
<td>Cancer laterality</td>
<td>30 (3)</td>
<td>0.577 (0.578)</td>
<td>0.825 (0.826)</td>
</tr>
<tr>
<td>Tumor size</td>
<td>374 (7)</td>
<td>0.721 (0.722)</td>
<td>0.788 (0.790)</td>
<td>0.756 (0.757)</td>
<td>Cancer laterality value</td>
<td>1241 (4)</td>
<td>0.993 (0.994)</td>
<td>0.991 (0.992)</td>
</tr>
<tr>
<td>Tumor size value</td>
<td>502 (305)</td>
<td>0.343 (0.344)</td>
<td>0.556 (0.557)</td>
<td>0.715 (0.716)</td>
<td>Cancer stage</td>
<td>109 (4)</td>
<td>0.923 (0.924)</td>
<td>0.967 (0.968)</td>
</tr>
<tr>
<td>Tumor site</td>
<td>44 (3)</td>
<td>0.021 (0.022)</td>
<td>0.817 (0.818)</td>
<td>0.889 (0.890)</td>
<td>Cancer stage value</td>
<td>131 (28)</td>
<td>0.835 (0.836)</td>
<td>0.767 (0.768)</td>
</tr>
<tr>
<td>Tumor site value</td>
<td>303 (173)</td>
<td>0.410 (0.411)</td>
<td>0.425 (0.426)</td>
<td>0.674 (0.675)</td>
<td>Histological type</td>
<td>49 (2)</td>
<td>0 (0)</td>
<td>0.934 (0.935)</td>
</tr>
<tr>
<td>Cancer grade</td>
<td>269 (5)</td>
<td>0.962 (0.963)</td>
<td>0.983 (0.984)</td>
<td>0.913 (0.914)</td>
<td>Histological type value</td>
<td>1009 (95)</td>
<td>0.917 (0.918)</td>
<td>0.903 (0.904)</td>
</tr>
</tbody>
</table>

Reference

Comparison of Natural Language Processing Approaches Identifying Opioid Overdose from Clinical Notes for Emergency Department

Vivienne Zhu, MD,1 Jingqi Wang, PhD,4 Jenna McCauley, PhD,2 Lindsey Jennings, MD,3 Hua Xu, PhD,4 Kathleen T. Brady, MD, PhD,2 Leslie A Lenert, MD, MS1

1Biomedical Informatics Center, 2Department of Psychiatry & Behavioral Sciences, 3Emergency Medicine at Medical University of South Carolina, Charleston, South Carolina, and 4School of Biomedical Informatics, University of Texas, Houston, Texas

Background. Over the last ten years, the number of individuals presenting to the emergency department (ED) as a result of intentional or unintentional opioid-related overdose (OD) has dramatically increased in the US. We developed and evaluated natural language processing (NLP) approaches for OD case identification and clinical characterization to support research and clinical practices.

Methods. The study cohort includes patients who were 12 years or older and visited the Medical University of South Carolina (MUSC) ED between 01-01-2013 and 12-31-2019. Patients with cancer or surgical visits were excluded. Selected clinical note types include ED note, ED provider note, ED nurse note, ED follow-up, progress note, discharge summary, and consultant note. Four NLP approaches were developed: 1) Named Entity Recognition (NER) + Rules: Leveraging CLAMP’s built-in default function modules (i.e., sentence detection, tokenizer, POS tagging, named entity recognition, assertion), we developed customized pipelines to extract information relevant to OD events. The customized OD dictionary consists of opioid medications (e.g., “narcotic,” “opiate,” “heroin,” “OXY,” “fentanyl,” “Oxycodone”), terms representing overdose (e.g., “overdose,” “OD,” “poison”), treatment for OD (e.g., “Naloxone,” “Narcan”), and patient’s positive response to Naloxone/Narcan (e.g., “reverse,” “respond,” “improve in mental status,” “awake”). We also included ICD-9/10 codes for OD (e.g., “T40.2X5A,” “T40.601A), as well as the descriptions of ICD-9/10 codes (e.g., “adverse effect of unspecified narcotics,” “poisoning by other opioids”) in the OD dictionary. After mining a small sample of data (clinical notes from 1,000 patients), we developed initial Apache Ruta rules to identify candidate OD cases at the sentence level when any one of the following criteria met: 1) mention of opioid medication and overdose irrespective of their order; 2) mention of Naloxone/Narcan administration and patient’s positive response to Naloxone/Narcan regardless of their order; or 3) mention of OD ICD-9/10 codes or its description. Ruta rules also excluded false positives. This approach also generated candidate sentences for other three approaches; 2) A Support Vector Machine (SVM) classifier based on java libsvm library, with unigrams, bigrams, and tri-grams as feature; 3) A Neural Network approach based on pre-trained context embeddings: Bidirectional Encoder Representations from Transformers (BERT) was used for the classification task; and 4) BERT+ Rules. Modifiers, such as OD for opioid, response for Narcan, subject, condition, hypothesis, or intentionality, were detected for two primary entities (opioid and Narcan). Inference rules were applied to further eliminate false positives, including situations, such as the subject is not the OD patient, an OD event did not happen (e.g., condition, hypothesis), or there is a negation (e.g., “no,” “deny”). Their performances were evaluated using the gold standard (manual review), and precision, recall, and F-score were reported.

Results. NER+ Rules identified 1,513 candidate OD sentences from the randomly sampled 2.47 million clinical notes. These sentences were annotated as True or False of OD case and then partitioned into three datasets: training (1,213), development (150), and test (150) with equally distributed True and False cases. In the test dataset, 114 true positive cases and 36 true negative cases (determined via expert chart review) were used to measure the performance of NLP approaches (Table 1).

Table 1. NLP performance comparison.

<table>
<thead>
<tr>
<th>NLP Approaches</th>
<th>Predicted</th>
<th>Correct</th>
<th>Precision</th>
<th>Recall</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>NER + Rules</td>
<td>137</td>
<td>105</td>
<td>0.77</td>
<td>0.92</td>
<td>0.84</td>
</tr>
<tr>
<td>SVM</td>
<td>122</td>
<td>105</td>
<td>0.86</td>
<td>0.92</td>
<td>0.89</td>
</tr>
<tr>
<td>BERT</td>
<td>117</td>
<td>103</td>
<td>0.88</td>
<td>0.90</td>
<td>0.89</td>
</tr>
<tr>
<td>BERT+Rules</td>
<td>106</td>
<td>103</td>
<td>0.97</td>
<td>0.90</td>
<td>0.94</td>
</tr>
</tbody>
</table>

Conclusion. NLP approaches could effectively extract OD information from ED clinical notes. The deep learning approach BERT with post-processing rules achieved the best performance.

Acknowledgment: This research was supported in part by the NIH grants (U01TR002628-01A1, K12DA031794-06A1, and R21LM012945) to MUSC.
Feasibility study of audio recording patient-clinician verbal communications in home healthcare settings
Maryam Zolnoori1, Sasha Vergez2, Zoran Kostic1, Siddhartha Reddy Jonnalagadda3, Maxim Topaz1
1Columbia University, 2Visiting Nurse Service of New York, 3Amazon Company

Background
Humans use speech as a communication tool to express their thought, intent, and emotion. Artificial intelligence methods, such as speech recognition, have the potential to capture and process speech. Emerging studies show that patients’ spontaneous speech in clinical encounters can provide indicators for mental disorders, patients’ emotional status, and cognitive impairment.1 However, currently, audio-recording of patients’ spontaneous speech in interaction with healthcare providers is not part of clinical workflows, and there is limited knowledge about the best practical approaches for audio-recording high quality patient-clinician verbal communication. This is particularly the case in home healthcare settings. This feasibility study aimed to identify the optimal audio-recording device that will be useful for audio-recording verbal communications of patient-clinician communication in home healthcare settings.

Method
This study was conducted at a large not-for-profit home health care organization located in New York, NY. The study was approved by an institutional review board at the organization. Initially, two members of our research team evaluated six voice recording devices (Black Vox 3652, SOTA Surveillance-USR5003, INSTAMIC PRO4, Sony ICD-TX6, Mini Wristband Voice Activated Recorder5, and Apple Watch) using criteria of ease of use, functionality (battery life and memory size), and quality of recorded voice in a controlled setting. Two recorders, Sony ICD-TX6 and Black Vox 365 were selected for further evaluation in home healthcare settings. Further, we worked with home healthcare nurses to record several patient-nurse encounters. We used the system usability scale (SUS) questionnaire and open-ended interviews to investigate nurses’ opinions about usability (ease of use and functionalities) of two selected recording devices. In the next step, two randomly selected audio files from each device were transcribed manually by our research assistant and automatically using the Amazon Medical transcribe system. We then used the word error rate (WER) to measure the quality of transcription by the AWS transcribe system. As a rule of thumb, the quality of the AWS transcribe system increases when there is an improvement of quality within the audio recorded data.

Results
We excluded the apple watch and INSTAMIC PRO in the first step of evaluation due to limited battery life (less than 2 days). Limited the battery life of recording devices might increase clinician workload because of the need for frequent charging. The ease of use and functionality of SOTA Surveillance-USR500 was similar to Black Vox 365, but the quality of recorded audio was lower, therefore this device was also excluded. We also excluded the Mini Wristband due to low quality of recorded audio. In the next step, the results of evaluation of Sony ICD-TX6 and Black Vox 365 in clinical setting showed that nurses prefer the Sony ICD-TX6 because of the inclusion of an indicator showing the status of recorders (on or off), sufficient battery life (up to 14 days) and good memory size (16 hours) for recording. Additionally, the total SUS score computed for Sony ICD-TX6 and Black Vox 365 were 97.5 and 42.5 respectively. The WER for Sony ICD-TX6 and Black Vox 365 were 0.36 and 0.53, indicating that the quality of automatic transcription by Amazon AWS service was higher for Sony ICD-TX6 compared with Vox 365.

Discussion and conclusion
Audio-recording of patient-clinician verbal communication in home healthcare settings is a challenging task. Selecting proper voice recording devices with high usability and capacity for recording high quality audio can improve the possibility of integrating audio-recording patient-clinician verbal communications into clinical workflows. Good quality of audio recording can improve the accuracy of downstream tasks built on analysis of this data, such as identifying patients with certain medical conditions such as depression. Our next steps will include further audio-recording of patient-nurse home healthcare encounters and extracting informative acoustic and linguistic features that can be used to analyze the patient’s voice and language for identifying signs (e.g., low speech rate) and symptoms (e.g., stress) associated with negative health outcomes (i.e., risk for rehospitalization).

References
SHARE-NW: An Innovative Public Health Informatics Tool

Uba Backonja, PhD, RN¹, Anne M. Turner, MD, MLIS, MPH¹, Betty Bekemeier, PhD, MPH, RN¹
¹Northwest Center for Public Health Practice, University of Washington, Seattle, WA

Introduction
Public health systems in the rural/frontier areas confront unique challenges improving health compared to urban areas. Rural communities have the highest rates of risky behaviors, the worst health outcomes of any U.S. population, and public health systems far more understaffed and poorly funded than urban areas.¹ Rural local health departments (LHDs) face barriers in identifying which populations face the greatest inequities to prioritize allocation of scarce resources. A major barrier is there are few rural-specific tools to easily access and analyze data, which is needed given rural LHDs often do not have personnel trained in data analytics.² To address this gap, we developed an innovative public health tool, Solutions in Health Analytics for Rural Equity across the Northwest (SHARE-NW). The aim is to support rural LHDs in the Northwest (Alaska, Washington, Idaho, and Oregon) by providing access to and training in using data to identify and address health inequities. The tool includes a web-based data dashboard and training resources. In this session we will describe our practice-centered approach, demonstrate features of the dashboard to identify inequities, and describe how our approach can be applied to a variety of public health issues. SHARE-NW is novel in applying user-centered design with rural public health staff and with a focus on equity.

The methods used to develop the SHARE-NW tool
We engaged end users (public health officials and health equity experts), through an iterative design process which included: a needs assessment completed in 2018,² paper mockup evaluations completed in 2019, and usability testing completed in 2020 and 2021. We identified publicly available datasets that included data on rural LHDs’ priority health outcomes identified during our needs assessment (e.g., state health departments, CDC, National Environmental Public Health Tracking Network Data, Feeding America); these data were used to generate visualizations. Due to the vast geographic distribution of our end-users, studies were conducted remotely using phone and teleconferencing software (Adobe Connect, Zoom). We developed the tool to include (1) visualizations to support data understanding and (2) trainings to support data use. To develop visualizations of publicly available data, we used Tableau and applied research on visualizing quantitative data and Munzner’s nested model for visualization design.² For curated trainings, we vetted existing trainings available online, applying the Public Health Learning Network’s Quality Standards for Training Design and Delivery Tool (mphli.org/phln/). We also developed trainings to address identified training gaps.

The resulting SHARE-NW tool
The SHARE-NW tool (available at nwcphp.org) showcases a dashboard that provides interactive visualizations of rural public health data. The dashboard includes data at the county level and multi-county level depending on each state’s unique way of organizing LHDs. There are dashboards for demographic data and health indicators organized by LHD health priority area. Users use the dashboard to view health outcomes and inequities in their county to compare their health jurisdictions with others in the 4-state catchment area. Users can download images of the graphs, download data, and view health indicator data by social determinants. The tool includes high quality trainings to enhance skills in health equity and social determinants, health priorities common to rural jurisdictions, and data-driven decision-making. The tool will be evaluated after launch to assess use, utility, and impact at the individual and LHD level.

Conclusion
SHARE-NW is an innovative informatics tool addressing a critical need among rural health practitioners to identify and address health equity. The session will demonstrate key features developed with end-users. SHARE-NW is funded by the Department of Health and Human Services Office of Minority Health (5 CPIMP171144-02-00).

References
The predominant model for data capture in multicenter clinical trials is for clinical research coordinators (CRCs) to abstract data from Electronic Health Records (EHR) and manually complete structured electronic case report forms. It is inefficient and usually retrospective; resulting in delays to availability of data, costly manual reviews, and risk of inaccurate/incomplete results. Greater utilization of eSource integrations can lead to enhanced site experiences, more efficient monitoring, more rapid identification of safety signals, and increased data integrity and quality.

OneSource reduces documentation burden by automating capture of regulatory-grade structured data, and supporting streamlined workflows that are integrated with healthcare delivery systems. Its implementation in the I-SPY COVID adaptive platform trial aids increased trial participation for critically ill COVID-19 patients in busy ICUs. A scalable design, built on the Substitutable Medical Apps – Reusable Technology on Fast Healthcare Interoperability Resources (SMART on FHIR) standard and on a general-purpose, configurable electronic data capture (EDC) platform (OpenClinica), allows use across multiple trials and sites. We will demonstrate user-centric EHR-integrated workflows and automated data acquisition that minimizes many of the inefficiencies and quality risks of today’s ‘swivel chair interoperability’ practices. It automatically logs users in and launches the proper EDC participant record based on the active EHR patient. Structured data from the EHR populates the eCRF, with workflows for user review/validation. eCRFs that cannot be directly populated are accessible for entry. Mobile support enables direct data capture at the point of care. The EDC’s proven 21 CFR part 11 compliant features support regulatory-grade evidence of data integrity.

Despite this significant level of integration, OneSource is deployable across heterogeneous sites and trials. Many eSource approaches rely on locally installed middleware and/or difficult-to-audit system-level APIs. OneSource uses a recognized security model, requires no software installs, and implements a user-driven and audited model for data transfers. This reduces burden and risk on hospital IT and security staff, currently a major barrier to eSource. The EDC’s study design tools allow mid-study eCRF and data mapping updates with audited publishing and change control - ensuring data integrity without involving site research or IT staff. Streamlined change management is a goal in all trials, and essential in a multi-arm adaptive platform trial like I-SPY COVID. The app is published in the Epic App Orchard and user testing has been completed in the test EHR environment at UCSF. Production rollout in I-SPY COVID will begin April 2021 at UCSF and scale to other Epic sites, followed by sites using Cerner and to other studies.

This work was supported, in part, by the Biomedical Advanced Research and Development Authority (BARDA), part of HHS within the office of the Assistant Secretary for Preparedness and Response, and Joint Program Executive Office, a part of the Department of Defense, under the Medical Chemical, Biological, Radiological, and Nuclear (CBRN) Defense Consortium (MCDC).

References
Histocartography: a Pipeline for Histology Image Analysis

Antonio Foncubierta-Rodríguez, PhD.\textsuperscript{1}, Pushpak Pati, MSc.\textsuperscript{1,2}, Guillaume Jaume, MSc.\textsuperscript{1,3}, Maria Gabrani, PhD.\textsuperscript{1}

\textsuperscript{1}IBM Research Europe, Rüschlikon, Switzerland; \textsuperscript{2}ETH Zurich, Zurich, Switzerland; \textsuperscript{3}EPFL, Lausanne, Switzerland

System purpose and description

An increasing number of pathology labs are moving to digital scanning and on-screen slide visualization. At the same time, machine learning promises to automatically perform time-consuming and complex tasks, from biomarker quantification or mitotic count to grading or subtyping. However, these promises come at the cost of an interpretability gap for pathologists: tissue samples are abstracted to pixels, analyzed by non-interpretable machine learning techniques like CNNs. We propose to reduce this gap with a web service that uses a representation of tissue samples based on biologically relevant entities, going back from pixels to cells and tissue types; and the use of an explainability method to quantify relevant concepts for the entities that were identified as the most important for the decision.

We propose to demonstrate the use of Histocartography for breast cancer lesion classification, including atypical cases, using HACT-Net\textsuperscript{2} and explain the decisions using a set of pathology-relevant concepts such as nucleus size, nucleus shape variation or chromaticity\textsuperscript{2} for the most important nodes. Evaluation of the results by a limited number of pathologists has shown that the cells identified as most discriminative also correlate with prior knowledge: e.g., the most important cells for malignant lesions have larger nuclei than those that are important for benign tissue.

Deployment degree

At the date of submission, the models are fully trained and the results can be accessed through the provided references. A REST API prototype is already deployed and an user-friendly interface is being developed. For developers, Histocartography has been released as an open source Python library at \url{https://github.com/histocartography/}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{flowchart.png}
\caption{(a) A region of interest is uploaded and processed, results are available in various formats. (b) Color-coded nuclei importance.}
\end{figure}

\textbf{Figure 1:} Visual representation of the system demonstration pipeline together with an example of output explanation.

References


Generalizing Dynamic Dashboards for Community-Engaged Interventions

Daniel R. Harris, Ph.D1,2, Daniel T. Redmond2, Jeffery C. Talbert, Ph.D2,3
1Institute for Pharmaceutical Outcomes and Policy, University of Kentucky, Lexington, KY.
2Center for Clinical and Translational Sciences, University of Kentucky, Lexington, KY.
3Institute for Biomedical Informatics, University of Kentucky, Lexington, KY.

Abstract
We present our web-based system that allows communities to configure, create, and customize visualizations based on their needs and interests when participating in a data-driven community-engaged intervention. The Communities That Heal (CTH) intervention was developed by the HEALing Communities Study (HCS) to significantly reduce opioid overdose deaths in 67 communities across four states. We leveraged open-source technologies to generalize a framework for dashboarding which enabled communities in the first wave of the intervention to select and customize visualizations most relevant for monitoring the opioid crisis in their community.

Introduction
The Communities That Heal (CTH) intervention required keeping community coalition leaders and members up to date on the local impact of the opioid overdose epidemic; by sharing community-specific dashboards, communities were able to have a localized view of measures and findings specific to the study. Communities were given an initial dashboard containing five high priority measures, such as the number of resident opioid overdose deaths or the number of high-risk opioid prescriptions. Communities were able to select additional measures defined by the study or to suggest new measures; each community independently explored topics most relevant to their own needs. For example, if a community wishes to link more individuals treated for overdoses in their emergency departments (ED) to medication for opioid use disorder, then they would focus on ED-related evidence-based practices and data; a community without an ED could focus on something more meaningful to their community.

Methods and Discussion
We organize data in accordance with pivotal elements of the intervention itself (a menu of evidence-based practices). Our web-based system allows users to interactively explore their data by offering customizable views and side-by-side comparisons. Users can dynamically filter by year, filter by reporting interval (monthly, quarterly, yearly), control graph types, size, and switch between raw counts and rates per population. A graph builder allows the creation of a customized graph containing data from several sources. We focused on open-source technologies to minimize the cost of ownership and to eliminate financial barriers for communities to adopt and continue the dashboards after the study concludes. In 2020, we had 164 users from 8 different communities. We are currently analyzing site usage metrics and usability to further improve our system; we wish to open-source our solution so other studies may benefit.

Figure 1. An interactive graph builder merges and visualizes different data sets based on user selections.

Acknowledgement
Research was supported by the National Institute on Drug Abuse of the National Institutes of Health under Award Number UM1DA049406. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References
Integrating Informatics for Integrating Biology and the Bedside with tranSMART: Flexible Data Warehousing with Complex Analytics

Jeffrey G. Klann, PhD; Michael Mendis; Peter M. Rice; Rudy Potenzone, PhD; Louisa May Klann; Griffin M. Weber, MD, PhD; Diane Keogh; Shawn N. Murphy, MD, PhD; Harvard Medical School; Partners Healthcare; Massachusetts General Hospital; Beth Israel Deaconess Medical Center; i2b2-tranSMART Foundation – Boston, MA; Oryza Bioinformatics Ltd, Royston, UK

Informatics for Integrating Biology and the Bedside (i2b2) is a well-established open-source clinical data warehousing and analytics platform in use at over 200 locations worldwide. [1] It was first released in November 2007, and over time it has evolved from a single-site cohort identification tool into an enterprise-ready, network-enabled analytics platform with a diverse open-science community. Originally funded by the National Institutes of Health, it is now maintained by the i2b2-tranSMART foundation, which continues to refine the core software according to users’ needs.

i2b2 serves a variety of user needs across the enterprise. i2b2 is an ideal cohort finding tool, frequently used as a self-serve clinical research portal for investigators. Its flexible data model makes it ideal for data ingestion and integration across a variety of source systems (from Electronic Health Record - EHR - to laboratory to gene expression data), and its application programming interface (API) allows it to serve as the backbone for large clinical research networks like the Clinical and Translational Science Award’s (CTSA’s) 50-site Accrual to Clinical Trials (ACT) network.

TranSMART was developed in 2008 by Johnson & Johnson from i2b2 to load and analyze data from clinical trials, including gene expression and other ‘omics’ datasets. They saw a lack of translatability of preclinical models into meaningful biological knowledge, so they extended i2b2 to support translational research studies by adding R-based data exploration and advanced analytics (e.g., correlation analysis, heat maps, principle component analysis, etc.). Clinical data and expression/proteomics/metabolomics and other data can be analyzed with a set of 20+ R-based workflows or exported as subsets to pass to local pipelines. The underlying database used i2b2 schemas for clinical data with new schemas for omics data, biomarkers, searches, and study metadata. tranSMART is used worldwide for translational research. For example, in the Human Hereditary Health in Africa Consortium (https://h3africa.org/), coordinators visit African villages and are able to input patient’s genomic information directly into tranSMART, for analyses of Chronic Kidney Disease’s genetic markers in African populations.

While i2b2 loads clinical data for all patients under a common terminology tree, tranSMART loads each study under an upper-level ‘study’ node with demographics, lab results, etc. as study-specific nodes below. Studies are not limited to human subjects: they can be for model organisms (e.g. mouse), cell lines or organoids.

Over time, the data model of tranSMART diverged, and the application software was never integrated with i2b2. Therefore, there is a need to harmonize the two tools so that they are compatible and interoperable. This would provide i2b2 users with a set of powerful R analytics tools, enabling data analytics across entire enterprise-wide i2b2 data repositories, and it would streamline the pipeline between cohort selection and data analysis.

The Foundation is engaged in several efforts to harmonize its two leading tools, so they can both be run on the same data. We developed the i2b2 Common Data Model, largely based on i2b2 version 1.7.12, with modifications to tranSMART 19 to use the same set of database tables and fields. The lack of guidance in the past on how to best represent patient data in i2b2 is part of the reason why i2b2 and tranSMART lost interoperability over time. The new Common Data Model documentation provides both descriptions of the core tables as well as information on how to use the data model with the i2b2 ontology, best practices and examples for different data types, and tips for extending the core data model. The i2b2 Common Data Model Guide is available at https://community.i2b2.org/wiki/display/i2b2/i2b2+Documentation.

Next, we aligned the tranSMART and i2b2 applications, so that the platforms can leverage each other’s strengths for both broad cohort-finding and complex genomic-enhanced data analyses. The latest tranSMART release used the full i2b2 data model, with additional tables added. The common schemas are now identical, and tranSMART creates all the additional i2b2 schemas and procedures to allow the two platforms to run on the same database. We are currently enhancing tranSMART so it can explore i2b2 native data as a single study.

Here we will demonstrate i2b2, tranSMART, and their integration, running on a single data set. The Common Data Model work is complete at the time of this writing. The application modifications to tranSMART to explore and analyze i2b2 datasets will be completed before the Symposium.
Interactive Model Report Card for Visual Exploration of Performance Heterogeneity and Biases on Population Subgroups

Bum Chul Kwon PhD¹, Uri Kartoun PhD¹, Shaan Khurshid MD²,³, Amit V Khera MD MSc²,³, Patrick T Ellinor MD PhD²,⁴, Steven A Lubitz MD MPH²,⁴, Kenney Ng PhD¹

¹Center for Computational Health, IBM Research, Cambridge, MA, USA; ²Cardiovascular Disease Initiative, Broad Institute of the Massachusetts Institute of Technology and Harvard University, Cambridge, MA, USA; ³Division of Cardiology, Massachusetts General Hospital, Boston, MA, USA; ⁴Cardiac Arrhythmia Service, Massachusetts General Hospital, Boston, MA, USA.

It is critically important for clinical researchers and practitioners to understand the differences and limitations of machine learning models within various subgroups of a population. To address this unmet need, we have developed an interactive visualization system to facilitate the identification of possible performance heterogeneity and biases. Recently, researchers investigated the concept of a model report card which summarizes how a model was trained and how it performs on a fixed set of subgroups in a static report (1,2). These approaches are significant milestones; however, they do not allow users to apply the model to local patient data sets, define custom subgroups, and explore how the model behaves within the subgroups.

The goal of the current system is to allow clinical researchers and healthcare data analysts to apply existing risk models to local patient data sets, visually inspect the performance and fairness of models within subgroups defined by key patient characteristics, such as comorbidities, age, sex, and race. The system includes multiple coordinated, and interactive visualizations so that users can freely define subgroups, explore model performance and fairness within and across subgroups, and understand how such metrics are associated with patients' clinical and demographic attributes. The system (Figure 1) is developed as a web-based prototype.

Figure 1. The design of interactive model report. Users can select a risk model and a local patient data set to launch the system. In the system, (a) users define subgroups by choosing comorbidities and/or demographic attributes; (b) users explore model performance (e.g., concordance index and calibration slope) and fairness measures (e.g., true positive rate differences and statistical parity differences); (c) users inspect features affecting model behavior using SHAP (SHapley Additive exPlanations) values and feature distributions.

References


The SMART Cumulus Text-to-FHIR NLP Pipeline

Timothy A Miller, PhD1,2, Bin Mao, PhD1, Daniel Gottlieb, MPA1,2, Kenneth Mandl, PhD1,2
1Boston Children’s Hospital, Boston, MA; 2Harvard Medical School, Boston, MA

Introduction

Unstructured text represents a large chunk of the important information in electronic health records, but is difficult to compute over due to the complexity of human language. Natural language processing (NLP) can convert unstructured text into structured texts, extracting references to diagnoses, signs and symptoms, procedures, and relations between these elements. We demonstrate implementations that map clinical notes directly to FHIR (Fast Healthcare Interoperability Resources) and compare two open-source NLP tools, Apache cTAKES and MedSpacy.

Background

FHIR has become the de facto standard for interchange of data between electronic health record (EHR) systems. The recent FHIR Bulk Data Standard extends the FHIR standard into population health, with application programming interface (API) hooks that support retrieval of large FHIR data sets. Regulations requiring EHR vendors to implement FHIR APIs portend a future where EHR data can be easily brought into analytics platforms for streamlined data analysis. However, these regulations do not resolve the issue of how to transform the unstructured content of clinical notes into FHIR resources. NLP tools exist for extracting concepts from notes and various other tasks, but modeling NLP outputs in FHIR is an unresolved problem. Existing mappings are closed source, use too-general resource types (Apache cTAKES fhir module) or do not track provenance of concepts (NLP2FHIR).

Implementation

We have developed mappings that convert outputs from broad semantic groups in the Unified Medical Language System (UMLS) to FHIR resources, and designed FHIR extensions to capture NLP-specific information. Procedures are mapped to the Procedure Resource, Signs/Symptoms are mapped to Observation Resources, Diseases/Disorders are mapped to Condition Resources, and Medications are mapped to MedicationStatement Resources. The extensions include versioning information for all processing software, document identifiers, and character offsets into the text where the concept was found. One of the primary goals of our implementation is that the automatic and less-certain nature of NLP-extracted resources is highlighted, and that provenance information is preserved so that downstream users can manually validate if required for their use cases. The existence of the modifier extension serves as the flag that the information is NLP-derived, and the document ID and concept span information serve as the link back to the provenance of the resource in the text.

Status of Implementation and Presentation Plans

We have developed two NLP systems that can both use the same NLP-to-FHIR mappings for their output, Apache cTAKES and MedSpacy. cTAKES is one of the most widely used open-source clinical NLP tools, and has a large existing userbase. MedSpacy is a relatively new tool that extends the widely used python-based Spacy NLP system to handle clinical text. It includes clinical specific sentence segmentation and a UMLS dictionary lookup similar to that in cTAKES. cTAKES has an existing REST server implementation, with an ad hoc json-formatted output string. Our new implementation maps from that output string to our proposed FHIR representations. We modified MedSpacy output to resemble the cTAKES REST output so that we can use the same mapping code to create FHIR resources from MedSpacy output as well. The cTAKES implementation is wrapped in a Docker container for ease of deployment, while work on containerizing the MedSpacy implementation is expected to finish shortly.

We first give a high-level overview of the mapping of NLP outputs to FHIR that we have proposed and implemented. We then show running versions of our container implementations, demonstrating how they can be queried with REST from standard browser tools and showing the FHIR output. We will show a pipeline on synthetic data that includes both structured and unstructured components, and show how our tool integrates the structured data with the FHIR output from the NLP systems, including sample queries that highlight use cases enabled by this framework.

Conclusion

We have developed a schema for representing information extracted from EHR notes into FHIR resources, and have created two implementations for open-source clinical NLP tools. The software will be made available open-source on GitHub.
Is your clinical decision support moving the needle on outcomes that matter?
Novel software for evaluating quality improvement initiatives

Evan Orenstein, MD1, Naveen Muthu, MD2, Marc Tobias, MD3
1Children’s Healthcare of Atlanta, Atlanta, GA; 2Children’s Hospital of Philadelphia, Philadelphia, PA; 3Phrase Health, Philadelphia, PA

Description: We present a novel system for associating clinical decision support (CDS) interventions with their intended clinical outcomes. This system ultimately uses a singular dashboard with these interwoven data elements in order to support quality improvement efforts within a health system. To do this, the software architecture streams data for CDS intervention performance, patient cohorts, and process and outcome measures to a cloud-based data pipeline that populates a web application. Ultimately, the software aims to lower the barrier to evaluate the effectiveness of CDS, like alerts and order sets, on complex and evolving clinical workflows and processes.

Problem: Although CDS has the potential to improve clinical outcomes, their actual impact has been inconsistent.1 Continuous evaluation is critical to achieve consistent improvement in outcomes that health systems, clinicians, and patients care about.5 However, most CDS initiatives are not assessed holistically due to limited resources and, when they are, many improvement efforts do not use rigorous methods to evaluate their processes.3 Existing approaches often use custom dashboards that require significant technical resources for each improvement initiative.

Innovation: Through funding from the National Library of Medicine, a multi-institutional research initiative coupled academic health systems and an industry partner to tackle issues related to data accessibility, data standardization, and human computer interaction. The resultant software utilizes a patent-pending extraction architecture to securely stream data from multiple data stores to a custom Amazon Web Services (AWS) data pipeline. System end-users can then apply the Institute for Healthcare Improvement’s model for improvement by constructing quality improvement projects with the data streams of choice and, with the help of directed questions, evaluate CDS effectiveness and receive recommendations for next steps.4

Adoption: The team embarked on user research in 2020 by interviewing over 30 informatics, quality, and IT stakeholders from two separate health systems. Following over 25 iterations, a live version is planned for implementation in Spring 2021 at 3 health systems participating in an implementation study of the platform.

Conclusion: Monitoring individual CDS elements is considered a best practice for electronic health record systems.5 However, CDS is invariably implemented to drive target downstream clinical processes and outcomes. Our system integrates data from a variety of sources and walks users through an evaluation framework to assess the impact of CDS on outcomes that are meaningful in patient care delivery rather than the individual CDS elements alone. Future efforts will aim to automate the effort to monitor process deviations and system changes.

References
VITAL: A semantic search engine for clinical text

Jingqi Wang, MS\textsuperscript{1}, Yaoyun Zhang, PhD\textsuperscript{1}

\textsuperscript{1}Melax Technologies, Inc; Houston, Texas, United States

Introduction

The last few decades see an explosion of unstructured clinical documents in electronic health records (EHRs). There is a great need for an easy-to-use tool that can effectively find useful information from EHR notes. Here we introduce VITAL, a newly developed semantic search engine for clinical text. By leveraging state-of-the-art information extraction technologies in the CLAMP natural language processing (NLP) toolkit\textsuperscript{1} and the clinical data standard of OMOP Common Data Model (CDM), VITAL provides a powerful yet user-friendly platform for quickly searching millions of clinical documents in EHRs, thus, to efficiently identify cohorts of patients of interest and their specific phenotypes. Use cases of VITAL include clinical chart review, healthcare quality reporting, disease registry curation, and clinical and translational research.

Overview:

VITAL is designed for efficient semantic search from clinical text, with configurable and scalable settings to support practical use cases. As illustrated in Figure 1, the workflow of VITAL includes several steps: data source preparation, NLP model deployment, information extraction and normalization, semantic indexing, query processing and search, and search result visualization.

Features of VITAL are as follows:

- **Customizable NLP models**: (1) adopt various existing NLP models following the UIMA framework (i.e., disorder attributes, medication-attributes, smoking status, breast cancer-recurrence etc.); (2) build new NLP pipelines (i.e., pathology notes processing) in CLAMP using lexicons, rules and state-of-the-art deep learning models. CLAMP built-in pipelines can be deployed to VITAL directly.

- **Data source management**: (1) support major relational databases including MySQL, PostgreSQL, Oracle, SQL server; (2) support user defined SQL to select and filtering on the input tables.

- **Multiple search modes**: (1) keyword-based search; (2) clinical concept-based search; (3) OMOP CDM-based query expansion; (4) Boolean search with combined queries (Figure 2).

- **Multi-view of search results**: (1) aggregate longitudinal data of the same patient (Figure 3); (2) highlight clinical variables in text with colors; (3) review patient lists to add tags and comments; (4) export results in the format of OMOP CDM table.

- **Task management**: (1) schedule and monitor automatic tasks of incremental processing and semantic indexing of new clinical documents; (2) support multiple common relational databases; (3) scalable architecture for processing millions of documents in parallel.

- **Flexible deployment**: (1) docker download with detailed implementation document on local servers; (2) cloud-based deployment based on AWS HIPAA compliance architecture.

Conclusion

Our studies have shown that Vital dramatically improves efficiency of manual chart review in both operational and research tasks. It will be freely available to the biomedical research community.

Reference: